# Report on Spatial Analysis of Global Variability in Covid-19

Parijat Chakraborty, Samarth Bhat, Soham Bakshi, Soham Ghosh

November 2020

## 1 Background

Coronavirus disease 2019 (COVID-19) is a contagious respiratory and vascular disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). First identified in Wuhan, China, it has caused an ongoing pandemic. The first human transmission of SARSCoV-2, the virus responsible for Covid-19, was reported to occur in Wuhan, China on November 17, 2019. On March 11, 2020, following identification of 118,000 Covid-19 cases in 114 countries responsible for nearly 4300 deaths, the World Health Organization recognized Covid-19 as a pandemic.

Ever since the onset of the pandemic, a number of online tools have become available to assist with the tracking of various statistical indicators related to global and local Covid-19 distribution. This project makes an attempt to relate various country parameters with Covid-19 parameters, and tries to analyse the dependence of Covid-19 incidence and prevalence, using various data available. This information may help to analyse the spatial distribution of global Covid-19 burden and help anticipate possible regions of Covid-19 outbreak.

## 2 Data Collection

In this project we collected data on country-specific variability in Covid-19 prevalence, incidence, and case fatality rate among 223 countries globally. We used the World Health Organization worldwide Covid-19 tracking site to determine the number of confirmed Covid-19 cases, the number of deaths attributed to Covid-19, and the case fatality rate for each of 223 countries. Using data from the United Nations Department of Economic and Social Affairs, we extracted key country-specific metrics with potential associations with Covid-19. We extracted country-specific economic, social and health related indicators from The World Bank Group Open Data database. All data were extracted on October 16, 2020. We developed a consolidated data set with the total of 18 indicators (including 5 Covid indicators) for 223 countries.

# 3 Brief Objectives

Using the consolidated data set with 13 country indicators and 5 Covid indicators, we undertake a descriptive and quantitative analysis of global impact of Covid 19 as follows:

1. Plot various Covid-19 indicators against various economic, social, and demographic indicators.
2. Study the plots using various statistical methods such as correlation and regression, with the aim to explain the observed relation of Covid-19 case incidence, case point prevalence, death incidence, death point prevalence and fatality.
3. Draw inference about global variance of point prevalence, incidence, mortality and fatality for Covid-19 cases.

# 4 Recreation of Data and Extension

This project is based on the paper "Spatial Analysis of Global Variability in Covid-19 Burden" by Miller L., Bhattacharyya R., Miller A. The paper had a data set of 238 countries with 7 Country parameters and 5 Covid parameters. The Covid data was accounted in the paper was for August 15, 2020. The paper inferred about independence of case incidence and case prevalence of Covid. However, no statistical methods were used to support their conclusion.

Our work of extension includes:

1. Creating an extended data set including 6 new country parameters like Diabetes prevalence, life expectancy, etc.
2. Recollecting the Covid data for October, 16th and updating the data set.
3. Recreating the 10 plots of the paper and creating all other possible plots of interested. (More than 100 plots were studied and the best fitted 65 plots are given)
4. Studying all the plots in details using methods like regression line, correlation coefficient and outlier analysis. Regression diagnostics are performed before using regression lines.
5. Making correlation plot of 13 Country indicators vs 5 Covid indicators to infer about their dependence.
6. Conducting chi-square test of hypothesis and $\beta_1 = 0$ test to understand median age vs fatality rate dependence.

# 5 Epidemiological terms

In this section we list and define some of the epidemiological terms that are used throughout the report:

1. **Point prevalence** refers to the prevalence measured at a particular point in time. It is the total number of persons with a particular attribute (mostly with some particular disease) on a particular date. The relevant point prevalence considered in this project is total Covid cases.

2. **Incidence** in epidemiology refers to the probability of occurrence of a given disease. Often it is calculated as number of new cases in some period of time. To keep the presentation consistent with the paper we shall consider incidence of Covid as total Covid cases per million population.

3. **Mortality rate** is generally taken as the total number of deaths caused by a particular disease divided by the concerned population. We shall measure it as deaths per million population caused by Covid.

4. **Case Fatality rate** is defined as the ratio of total deaths to total cases for a particular disease. The relevant case fatality rate is calculated as Covid deaths divided by Covid cases.

# 6 Global occurrence of Covid-19

The analysis included 38,711,983 Covid-19 cases (1,094,220 associated deaths), representing 99.66% of total global confirmed cases and 99.87% of total global confirmed Covid-19 deaths(collected from sample of 223 countries).

Covid-19 cases were reported in 208 (93.27%) out of the 223 countries considered, with 186 (83.41%) out of the 223 countries reporting at least one related death. In total, 216 countries have at least one Covid-19 case, out of which 191 countries reported at least one related death.

The following tables list the top 10 countries based on Total Cases, Cases per million population, Total deaths and Deaths per million population:

| Country | Total Cases | Country | Cases Per Million |
|---|---|---|---|
| United States | 7833851 | Bahrain | 45225.15904 |
| India | 7370468 | Qatar | 44706.91792 |
| Brazil | 5140863 | Andorra | 41286.48159 |
| Russia | 1369313 | Aruba | 39572.52309 |
| Argentina | 931967 | Israel | 34328.32286 |
| Colombia | 930159 | Colombia | 34260.51788 |
| Spain | 921374 | French Guiana | 32459.42572 |
| Peru | 856951 | Holy See | 28304.65701 |
| Mexico | 829396 | Panama | 26697.83502 |
| France | 780994 | Kuwait | 25990.37955 |

Table 1: Top 10 countries with highest incidence and prevalence of Covid 19 cases

| Country | Total Deaths | Country | Deaths Per Million |
|---|---|---|---|
| United States | 215199 | San Marino | 1237.806136 |
| Brazil | 151747 | Peru | 1016.382033 |
| India | 112161 | Belgium | 891.0557315 |
| Mexico | 84898 | Andorra | 763.6057723 |
| The United Kingdom | 43293 | Spain | 717.6378851 |
| Italy | 36372 | Bolivia | 717.6377049 |
| Spain | 33553 | Brazil | 713.9039152 |
| Peru | 33512 | Chile | 702.7546948 |
| France | 32868 | Ecuador | 697.4982903 |
| Iran | 29605 | Mexico | 658.4672864 |

Table 2: Top 10 countries with highest incidence and prevalence of Covid 19 deaths

# 7 Global Demographic Maps

In this section we attach the global demographic maps to help visualizations of spatial distribution of incidence and prevalence of Covid cases and deaths.

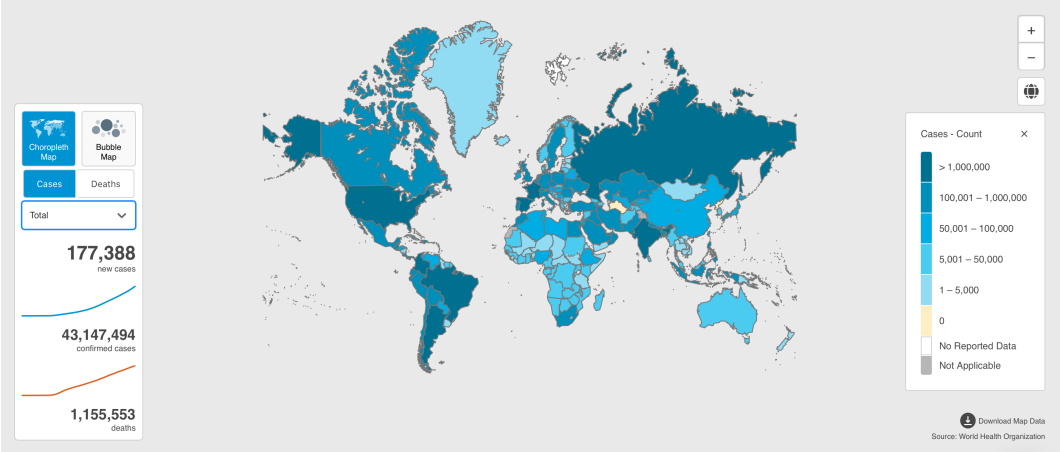



Figure 1: Global Demographic Map of Total Covid-19 Cases Distribution

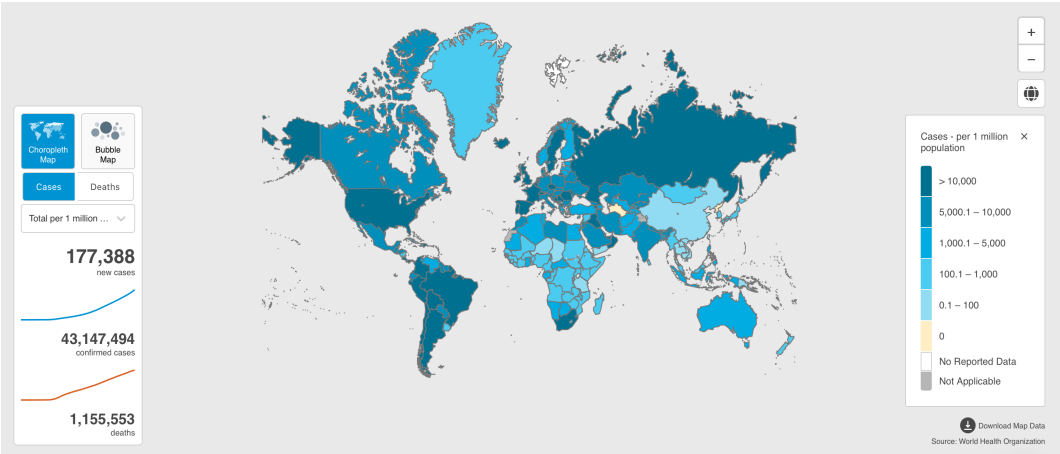



Figure 2: Global Demographic Map of Covid-19 Cases Per Million Distribution

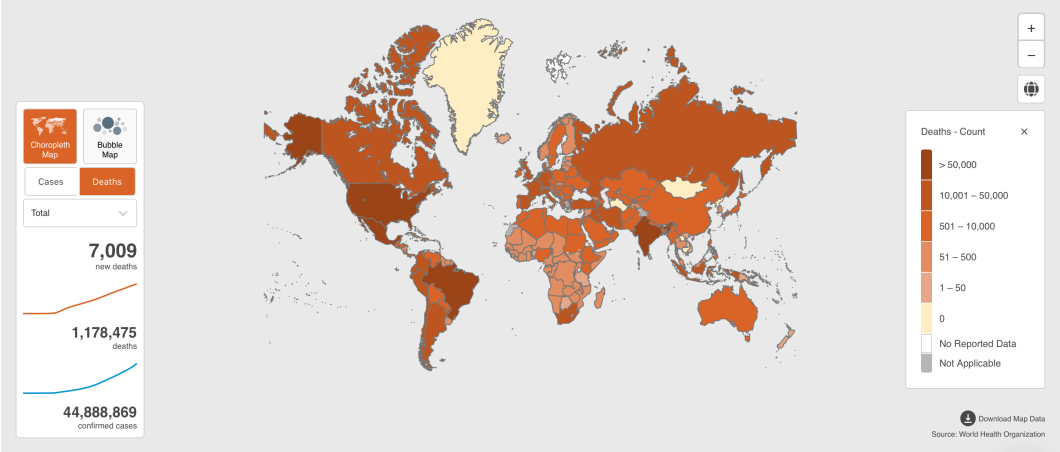



Figure 3: Global Demographic Map of Total Covid-19 Deaths Distribution

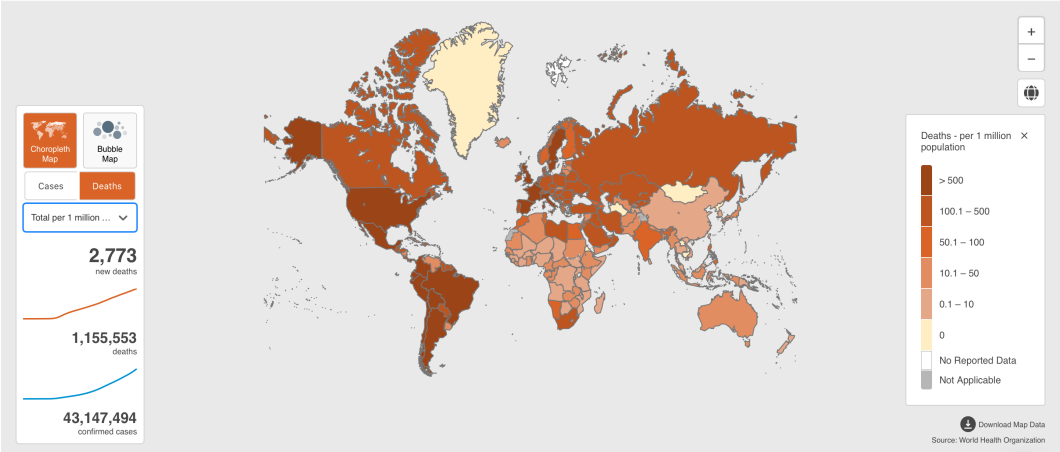



Figure 4: Global Demographic Map of Deaths Per Million Covid-19 Distribution

# 8 Statistical Methods

In this section we list and describe the statistical methods that have been utilized in this project:

## 8.1 Correlation

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a statistic that measures linear correlation between two variables $X$ and $Y$. It has a value between $+1$ and $-1$. Pearson's correlation coefficient, when applied to a population, is commonly represented by the Greek letter $\rho$ and may be referred to as the population correlation coefficient or the population Pearson correlation coefficient. Given a pair of random variables $(X, Y)$, the formula for $\rho$ is:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

where, $Cov$ is the covariance, $\sigma_X$ and $\sigma_Y$ are the standard deviations of $X$ and $Y$ respectively.

**Interpretations:** Correlations equal to $+1$ or $-1$ correspond to a bivariate distribution entirely supported on a line (in the case of the population correlation). A value of 1 implies that a linear equation describes the relationship between $X$ and $Y$ perfectly for which $Y$ increases as $X$ increases. A value of $-1$ implies that all data points lie on a line for which $Y$ decreases as $X$ increases. A value of 0 implies that there is no linear correlation between the variables.

Let $\overline{X}$, and $\overline{Y}$ be the means of $X$ and $Y$ respectively. More generally, note that $(X_i - \overline{X})(Y_i - \overline{Y})$ is positive if and only if $X_i$ and $Y_i$ lie on the same side of their respective means. Thus the correlation coefficient is positive if $X_i$ and $Y_i$ tend to be simultaneously greater than, or simultaneously less than, their respective means. The correlation coefficient is negative (anti-correlation) if $X_i$ and $Y_i$ tend to lie on opposite sides of their respective means. Moreover, the stronger is either tendency, the larger is the absolute value of the correlation coefficient.

## 8.2 Simple Linear Regression

In statistics, simple linear regression is a linear regression model with a single explanatory variable. That is, it concerns two-dimensional sample points with one independent variable and one dependent variable (conventionally, the $X$ and $Y$ coordinates in a Cartesian coordinate system) and finds a linear function (a non-vertical straight line) that, as accurately as possible, predicts the dependent variable values as a function of the independent variable. It is assumed that the original relation between $X$ and $Y$ is a linear, and is given by a equation

$Y = \beta_0 + \beta_1 X + E$, where $E$ is some random error.

In simple linear regression, the deterministic component of the probability model is given by a linear equation $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ to predict the actual response $Y_i$, where $\hat{\beta}_0$ and $\hat{beta}_1$ are estimates of $\beta_0$ and $\beta_1$. In this method, we make a prediction error (or residual error) of size $\epsilon_i = Y_i - \hat{Y}_i$.

A line that fits the data "best" will be one for which the $n$ prediction errors (where $n$ is the dataset size) — one for each observed data point — are as small as possible in some overall sense. One way to achieve this goal is to invoke the "least squares criterion," which says to "minimize the sum of the squared prediction errors." That is:
The equation of the **best fitting line** is: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

We just need to find the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that make the sum of the squared prediction errors the smallest it can be. That is, we need to find the values $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize:

$$Q = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$$

Using calculus, one can minimze $Q$ with respect to $\beta_0$ and $\beta_1$ to get:

$\hat{\beta}_1 = \frac{Cov(X,Y)}{\sigma_X^2}$ and $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

where the symbols have their usual meaning.
Now we state the assumptions needed to get a meaningful and accurate best fitting line using simple linear regression:

1. The plot $Y$ vs $X$ is more or less linear.
2. The mean of the probability distribution of the errors $E$ is 0.
3. The variance, $\sigma^2$, of the probability distribution of the errors $E$ is constant. (Homoscedasticity)
4. The probability distribution of the errors $E$ is normal.
5. The values of the errors $E_i$ associated with any two values of $Y_i$ are independent.

An alternative way to describe all four assumptions is that the errors, $E_i$, are independent normal random variables with mean zero and constant variance, $\sigma^2$. Note that the assumption of normality of the distribution of the error $E$ is needed only to perform inferential tests, making use of the normality of the distribution.

**Caution:**

**Errors** and **residuals** are two closely related and easily confused measures of the deviation of an observed value of an element of a statistical sample from its "theoretical value". The **error** (or disturbance) of an observed value is the deviation of the observed value from the (unobservable) true value of a quantity of interest (for example, a population mean), and the **residual** of an observed value is the difference between the observed value and the estimated value of the quantity of interest (for example, a sample mean).

## 8.3 Simple Linear Regression Diagnostics

In the previous part, we described how to do ordinary linear regression. Without verifying that the data have met the assumptions underlying OLS regression, results of regression analysis may be misleading.

Here will explore how you can use R to check on how well our data meet the assumptions of Ordinary Least Squares (OLS) regression. In addition to the assumptions listed above, there are issues that can arise during the analysis that, while strictly speaking are not assumptions of regression, are none the less, of great concern to data analysts. One of them is influence: individual observations that exert undue influence on the coefficients. We will not deal with this in our project.

R has many of the methods needed for the diagnostics in stats package which is already installed and loaded in R. There are some other tools in different packages that we can use by installing and loading those packages in our R environment.

Since error is theoretical, and cannot be observed for practical use. Thus in practice, statisticians use residual as an approximate of error, and perform regression diagnostic using residuals. In the following, since we deal with regression diagnostics, we replace "errors" by "residuals". However one must keep in mind that residuals most often cannot be truly normal, independent, homoscedastic or have mean exactly 0.

### 8.3.1 Checking mean of errors is zero

Note that the first condition that the mean of the probability distribution of the residuals $\epsilon$ is 0 is equivalent to the condition that the mean of the $Y_i$'s is linear with $X_i$.

To check linearity residuals should be plotted against the fit as well as other predictors. If any of these plots show systematic shapes, then the linear model is not appropriate and some nonlinear terms may need to be added. In package car, function residualPlots() produces those plots. It also gives a test of the appropriateness of linear model by adding quadratic term for each variable (testing for curvature).

### 8.3.2 Checking homoscedasticity

One of the main assumptions for the ordinary least squares regression is the homogeneity of variance of the residuals. If the model is well-fitted, there should be no pattern to the residuals plotted against the fitted values. If the variance of the residuals is non-constant then the residual variance is said to be "heteroscedastic." There are graphical and non-graphical methods for detecting heteroscedasticity. A commonly used graphical method is to plot the residuals versus fitted (predicted) values.

### 8.3.3 Checking independence

A simple visual check would be to plot the residuals $\epsilon$ versus the explanatory variable $X$. If the model is well-fitted, there should be no pattern to the residuals plotted against the explanatory variables.

### 8.3.4 Checking normality

Normality of residuals is only required for valid hypothesis testing, that is, the normality assumption assures that the p-values for the t-tests and F-test will be valid. Normality is not required in order to obtain unbiased estimates of the regression coefficients. OLS regression merely requires that the residuals (errors) be identically and independently distributed. Furthermore, there is no assumption or requirement that the predictor variables be normally distributed.

Furthermore, because of large sample theory if we have large enough sample size we do not even need the residuals be normally distributed. However for small sample sizes the normality assumption is required.To test normality we use qq-normal plot of residuals.

As far as our project is concerned, since our sample is almost the whole population (thus a large sample), we will not be requiring normality of residuals.

### 8.3.5 Unusual and Influential data

A single observation that is substantially different from all other observations can make a large difference in the results of linear regression analysis. If a single observation (or small group of observations) substantially changes the results, we would want to know about this and investigate further. There are three ways that an observation can be unusual:

1. **Outliers**: In linear regression, an outlier is an observation with large residual. In other words, it is an observation whose response-variable value is unusual given its values on the explanatory variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.

2. **Leverage**: An observation with an extreme value on a explanatory variable is called a point with high leverage. Leverage is a measure of how far an observation deviates from the mean of that variable. These leverage points can have an effect on the estimate of regression coefficients.

3. **Influence**: An observation is said to be influential if removing the observation substantially changes the estimate of coefficients. Influence can be thought of as the product of leverage and outlierness.

**Identifying outliers:** Studentized residuals can be used to identify outliers. In R we use rstudent() function to compute Studentized residuals. One should pay attention to studentized residuals that exceed $+2$ or $-2$, and get even more concerned about residuals that exceed $+2.5$ or $-2.5$. For this report, and related analysis, we only give minimal details of the calculations behind the Studentized residuals, in the following "Theory" subsection.

**In this project, we do not deal with the identification of leverage and influential points.**

#### 8.3.6 Theory:

In the following we only deal with a minimal overview of the theory behind studentized residuals, which are used in regression analysis. We only deal with the theory when restricted to simple linear regression. The hat-value $h_i$ is a common measure of leverage in regression. These values are so named because it is possible to express the fitted values $\hat{Y}$ in terms of the observed values $Y_i$ :

$$\hat{Y_j} = \sum_{i=1}^{n} h_{ij} Y_i$$

1. Thus, the weight $h_{ij}$ captures the contribution of observation $Y_i$ to the fitted value $\hat{Y_j}$: If $h_{ij}$ is large, then the $i^{th}$ observation can have a substantial impact on the $j^{th}$ fitted value.

2. Properties of the hat-values:

   (a) $h_{ii} := \sum_{j=1}^{n} {h_{ij}}^2$, and so the hat-value $h_i \equiv h_{ii}$ summarizes the potential influence (the leverage) of $Y_i$ on all of the fitted values.

   (b) $\frac{1}{n} \leq h_i < 1$

   (c) In simple-regression analysis, the hat-values measure distance from the mean of X:

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{j=1}^{n} (X_j - \overline{X})^2}$$

**Detecting Outliers: Studentized Residuals**

Discrepant observations usually have large residuals, but even if the errors $E_i$ have equal variances (as assumed in the general linear model), the residuals $\epsilon_i$ do not:

$$V(\epsilon_i) = \sigma_E^2(1 - h_i)$$

Standardised residuals are defined as

$$\epsilon_i' = \frac{\epsilon}{\sqrt{V(\epsilon_i)(1 - h_i)}}$$

However since the variances of errors is often not available, and hence one estimates it as $S_\epsilon^2 = \sum_{j=1}^{n} \frac{\epsilon_j^2}{n-2} = \frac{SSE}{n-2}$.

We can then form a **internally studentized residual** by calculating

$$\epsilon_i" = \frac{\epsilon_i}{S_\epsilon\sqrt{1 - h_i}}$$

$\epsilon_i$" follows a $t$-distribution with $n - 2$ degrees of freedom. However, high-leverage observations tend to have small internally studentized residuals, because these observations can coerce the regression line to be close to them: When $|\epsilon_i|$ is large, $S_\epsilon^2 = \sum_{j=1}^{n} \frac{\epsilon_j^2}{n-2}$, which contains $\epsilon_i^2$, tends to be large as well.

Suppose that we refit the model deleting the $i^{th}$ observation, obtaining an estimate $S_\epsilon(-i)$ of $\sigma_E$ that is based on the remaining $n - 1$ observations, that is :

$$S_\epsilon(-i) = \frac{(\sum_{j=1}^{n} \epsilon_j^2) - \epsilon_i^2}{n - 3}$$

Then the **externally studentized residual** (or **jack-knife residual**) $\epsilon_i^\star$ is

$$\epsilon_i^\star = \frac{\epsilon_i}{S_\epsilon(-i)\sqrt{1 - h_i}}$$

has independent numerator and denominator, and follows a $t$- distribution with $n - 3$ degrees of freedom.

# 9 Detailed Study of Country Indicator - Covid Indicator Plots

In this section we analyze the scatter plots of Country indicators vs Covid indicators using statistical methods like regression and correlation. Raw data for the analysis is the compiled data set with total 18 indicators (including 5 Covid indicators) and 223 countries. Logarithmic plots are used when the indicators had large values, in cases when the indicators had 0 values $log(1 + indicator)$

```
#Data Sets


MedianAge=Compiled_Data[is.na(`Median Age years`)==F,]
lifeexpect=Compiled_Data[is.na(`Life Expectancy`)==F,]
pov=Compiled_Data[is.na(`Poverty Headcount ratio at 1.90USD a day by percentage of population`)==F,]
Population=Compiled_Data[is.na(`Population`)==F,]
PopulationDensity=Compiled_Data[is.na(`Population Density`)==F,]
LandArea=Compiled_Data[is.na(`Land Area`)==F,]
UrbanPopulation=Compiled_Data[is.na(`Urban Population`)==F,]
gdp.new=Compiled_Data[is.na(`GDP USD`)==F,]
pc.new=Compiled_Data[is.na(`Per Capita USD`)==F,]
lit.new=Compiled_Data[is.na(`Literacy`)==F,]
diabetes.new=Compiled_Data[is.na(`Diabetes`)==F,]
healthexp.new=Compiled_Data[is.na(`Health Expenditure per capita PPP USD`)==F,]
deathcomm.new=Compiled_Data[is.na(`Death by communicable disease or malnutrition in percentage`)==F,]
```

Figure 5: Definition of Datasets in R

is used. Correlation coefficient is calculated for the same functions of indicators used in the plot. Regression diagnostic tests are done for the plots looking somewhat linear. The diagnostic tests are attached in a supplementary file- only those are given for which the plots passed the regression diagnostic tests. For each Country indicator we study 5 plots corresponding to the 5 Covid indicators (Total Cases, Cases Per Million, Total Deaths, Deaths Per Million, Fatality Rate).

Figure 5 shows how the datasets were defined in R. For every country indicator considered, a separate dataset was made for the indicator, excluding the countries whose data for that indicator was not available. For a particular country indicator, any plot versus a COVID indicator was made using the dataset of that indicator.

## 9.1 Population

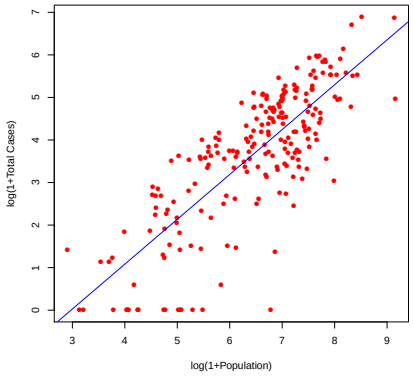
From the plot (a) it is evident that population and total cases increases with each other. The logarithmic indicators shows linear relation. The Pearson's correlation coefficient is calculated to be 0.80. The plot passes regression diagnostic test. The regression parameters are calculated to be

$$\hat{\beta_0} = -3.134 \text{ and } \hat{\beta_1} = 1.054.$$

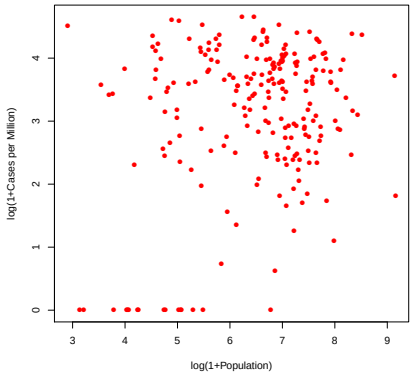
Hence, the regression line for $Y = log(1 + TotalCases)$ and $X = log(1 + Population)$ is given by

$$Y = 1.054X - 3.134.$$

On the other hand the Population vs Cases Per Million plot fails to show any linear tendency. The Pearson's correlation coefficient is calculated to be 0.23.

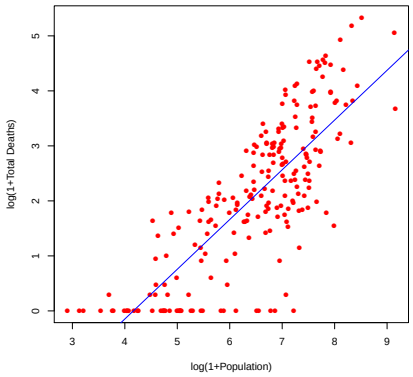



(a) Population vs Total Cases

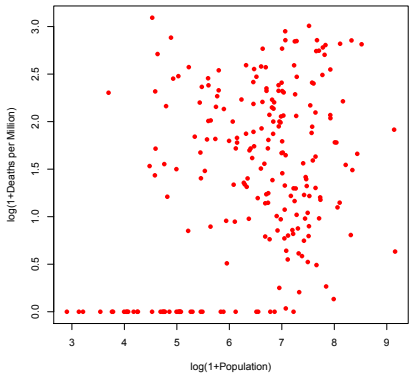



(b) Population vs Cases Per Million

Also we observe lack of linearity in the plot. Hence we don't use regression line for this plot.




(a) Population vs Total Deaths




(b) Population vs Deaths Per Million

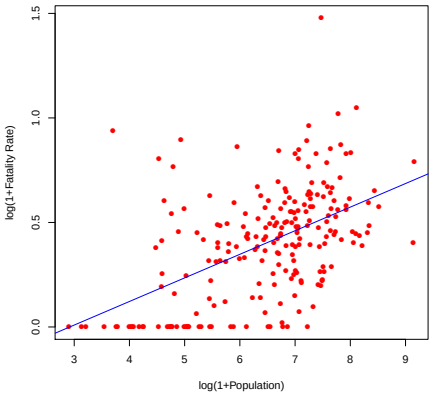
The plot for Population vs Total Deaths also exhibits linear tendency when when log-transformed indicators are taken. The appropriate Pearson's correlation coefficient is calculated to be 0.79. The plot passes regression diagnostic test. The regression parameters are

$$\hat{\beta}_0 = -3.7640 \text{ and } \hat{\beta}_1 = 0.9042.$$

Thus, the equation of regression is-

$$Y = 0.9042X - 3.7640$$

where $X = log(1 + Population)$ and $Y = log(1 + TotalDeaths)$.




Population vs Fatality Rate

Population vs Deaths Per Million plot is not showing any linear tendency. The correlation coefficient for the logarithmic indicators is 0.34.

The plot of Population vs Fatality Rate is a moderately linear plot. The Pearson's correlation coefficient of this plot is 0.51. This correlation is neither high nor low. The plot passes regression diagnostic test. The regression parameters are

$$\hat{\beta}_0 = -0.3294 \text{ and } \hat{\beta}_1 = 0.1128.$$

Hence, the regression line is

$$Y = 0.1128X - 0.3294$$

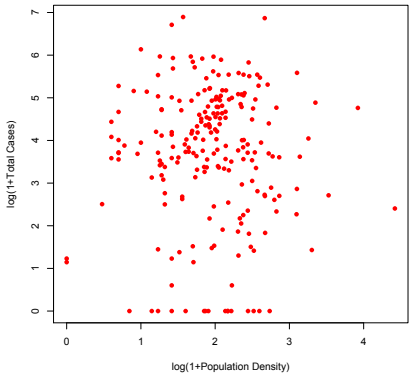
where $X = log(1 + Population)$ and $Y = log(1 + FatalityRate)$.

## 9.2 Population Density

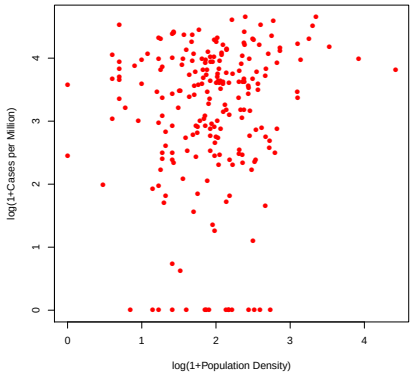
The Pearson's correlation coefficients of the plots of Population Density vs Covid indicators are comparatively low. The correlations for the plots of Population density vs Total Cases and Population Density vs Cases Per Million are $-0.02$ and 0.10 respectively.

The Pearson's correlations calculated for Population Density vs Total Deaths plot is -0.07 and for Population Density vs Deaths Per Million is 0.03. As the correlation values are too low and the plots does not look linear we don't use regression lines in the plots.

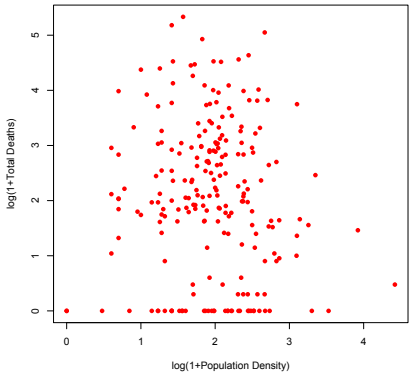
For the plot of Population Density vs Fatality Rate the correlation coefficient is -0.076. Also we cannot observe any linear relation between them from the plots.



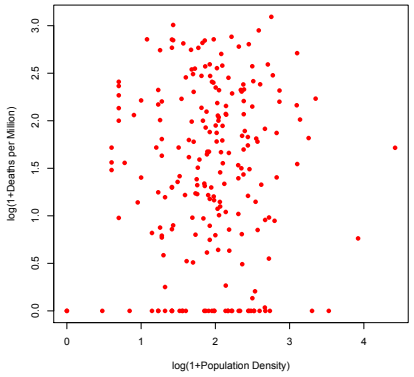

(a) Pop Density vs Total Cases




(b) Pop Density vs Cases Per Million




(a) Pop Density vs Total Deaths




(b) Pop Density vs Deaths Per Million

## 9.3 Gross Domestic Product (GDP)

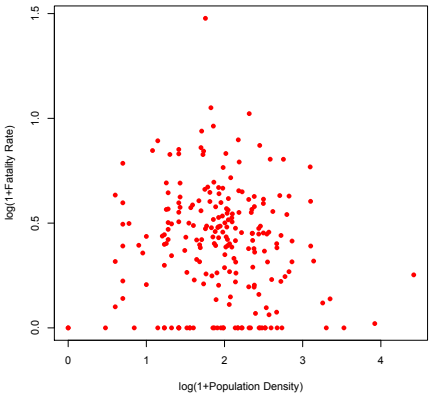
US dollar is taken as the unit of GDP. Firstly, we look at the plots of GDP vs Total Cases. From the plot itself it can be guessed that there is a linear relationship between $X = log(1 + GDP)$ and $Y = log(1 + TotalCases)$. The Pearson's correlation coefficient for the plot is calculated to be 0.82.

The plot of GDP vs Total Cases also looks quite linear and correlation is sufficiently high. The plot passes regression diagnostic test. The regression parameters are

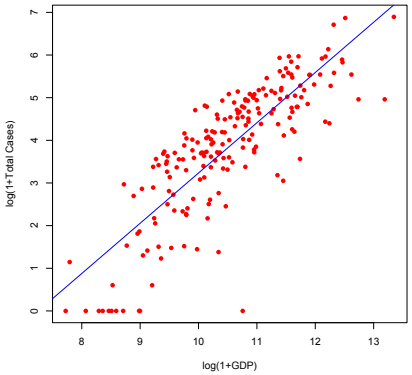
$$\hat{\beta}_0 = -8.548 \text{ and } \hat{\beta}_1 = 1.178.$$

Hence, the regression line is

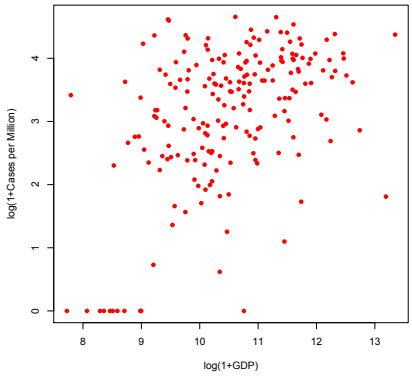
$$Y = 1.178X - 8.548.$$




Population Density vs Fatality Rate




(a) GDP vs Total Cases



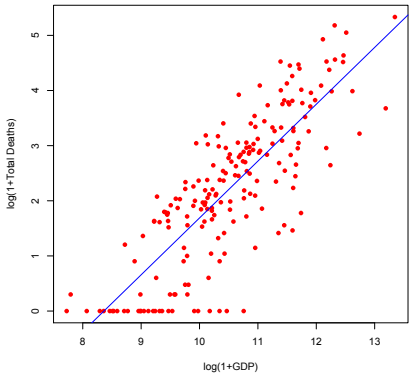

(b) GDP vs Cases Per Million

On contrary, the plot of GDP vs Cases Per Million is not showing any linear tendency. The correlation of the logarithmic indicators is calculated to be 0.46.

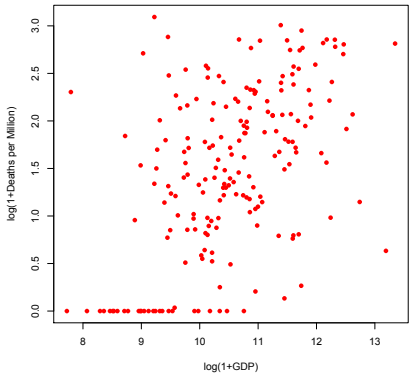
GDP vs Total Deaths is having linear tendency in the plot. The Pearson's correlation coefficient for the GDP vs Total Deaths Plot is 0.81. Due to the observable linear trend. The plot passes regression diagnostic test.

The regression parameters are

$$\hat{\beta}_0 = -8.622 \text{ and } \hat{\beta}_1 = 1.031.$$




(a) GDP vs Total Deaths



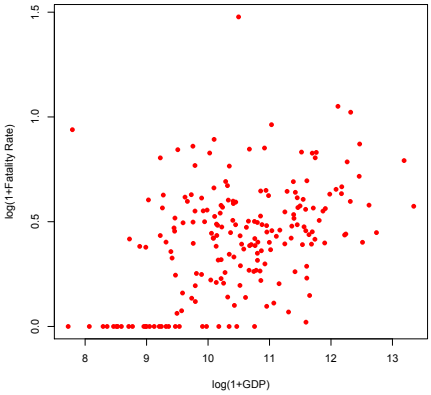

(b) GDP vs Deaths Per Million

So, equation of the regression line is

$$Y = 1.031X - 8.622$$

where $X = log(1 + GDP)$ and $Y = log(1 + TotalDeaths)$.

On the other hand, the plot of GDP vs Deaths Per Million is not showing much linear tendency. The correlation of the logarithmic indicators is calculated to be 0.49.

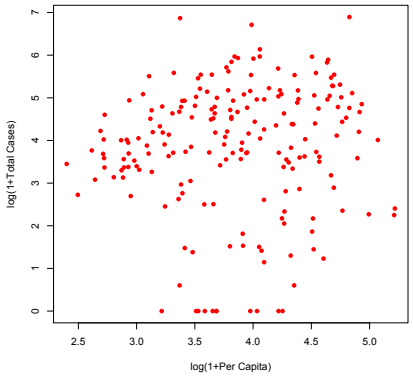



GDP vs Fatality Rate

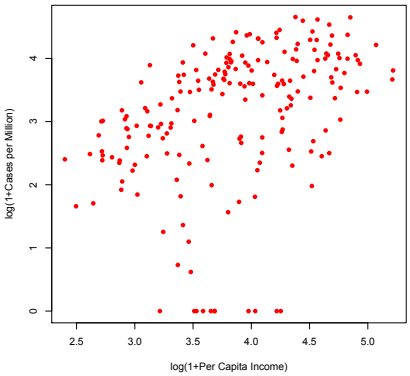
The Pearson's coefficient calculated for the plot of GDP vs Fatality Rate is 0.42. As the value is comparatively low we don't use regression line here.

## 9.4 Per Capita Income (PCI)

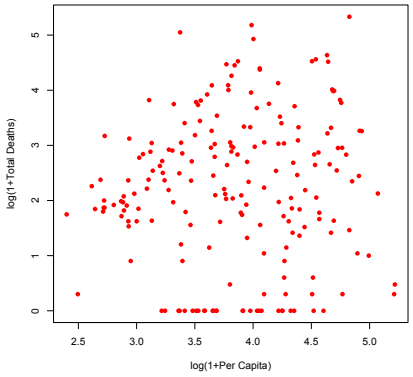
US dollar is taken as the unit of Per Capita Income. Firstly, we look at the plots of Per Capita Income vs Total Cases. The Pearson's correlation coefficients of the plots of Per Capita Income vs Covid indicators are comparatively low. The correlations for the plots of Per Capita Income vs Total Cases and vs Cases Per Million are 0.08 and 0.39 respectively. It is interesting to note that Per Capita Income has a much lower correlation with total cases, than cases per million.



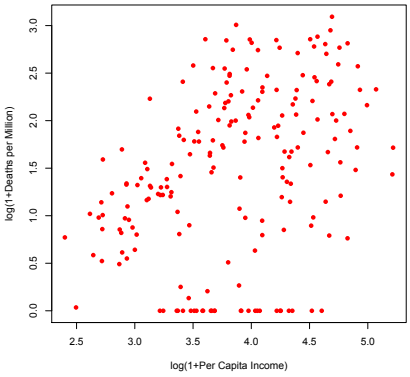

(a) PCI vs Total Cases




(b) PCI vs Cases Per Million

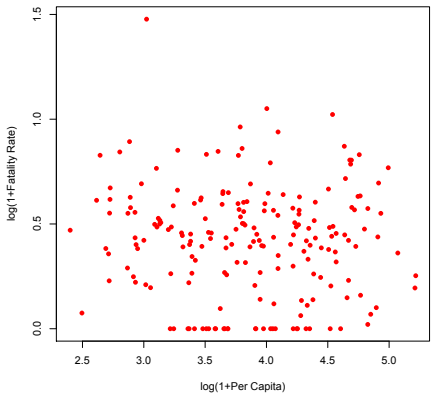



(a) PCI vs Total Deaths




(b) PCI vs Deaths Per Million

The Pearson's correlations calculated for Per Capita Income vs Total Deaths plot is 0.065 and for Per Capita Income vs Deaths Per Million is 0.37. Analogous to the plots of Per Capita Income vs Total cases and vs cases per million, we see that Per Capita Income has a much lower correlation with Total deaths than deaths per million.




Per Capita Income vs Fatality Rate

For the plot of Per Capita Income vs Fatality Rate the correlation coefficient is -0.062. Besides such a low coefficient value we can't conclude any linear relation between them from the plots as well.

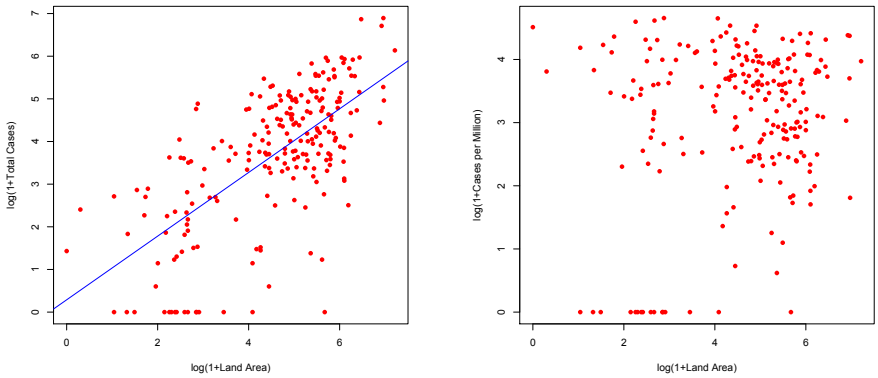
## 9.5 Land Area

The plot of Land Area vs Total Cases indicates a linear relationship between $X = log(1 + GDP)$ and $Y = log(1 + TotalCases)$. The Pearson's correlation coefficient for the plot is calculated to be 0.67. The correlation is sufficiently high, and the plot looks somewhat linear. The plot passes regression diagnostic test. We go for regression line. The regression parameters are

$$\hat{\beta}_0 = 0.2884 \text{ and } \hat{\beta}_1 = 0.7464.$$

Hence, the regression line is

$$Y = 0.7464X + 0.2884.$$




(a) Land Area vs Total Cases

(b) Land Area vs Cases Per Million

On contrary, the plot of Land Area vs Cases Per Million is not showing any prominent linear tendency. The correlation of the logarithmic indicators is calculated to be 0.15. The low correlation value further indicates the lack of linear tendency in the plot.

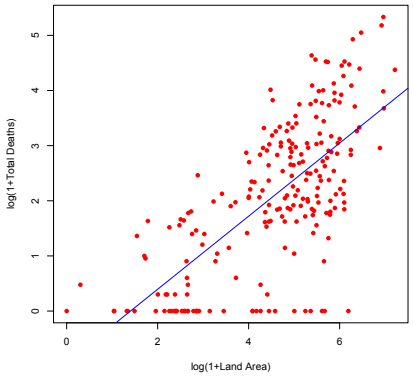
Land Area vs Total Deaths is having linear tendency in the plot. The Pearson's correlation coefficient for the Land Area vs Total Deaths Plot is 0.69. We see a linear trend and high correlation coefficient. The plot passes regression diagnostic test. The regression parameters are

$$\hat{\beta}_0 = -0.9301 \text{ and } \hat{\beta}_1 = 0.6628.$$

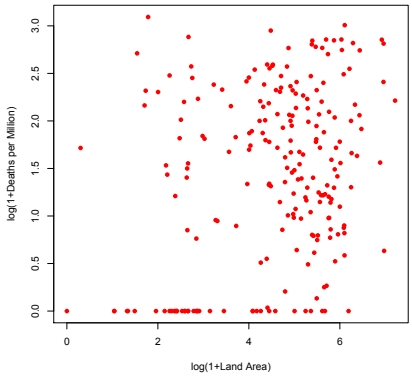
So, equation of the regression line is

$$Y = 0.6628X - 0.9301$$

where $X = log(1 + LandArea)$ and $Y = log(1 + TotalDeaths)$.

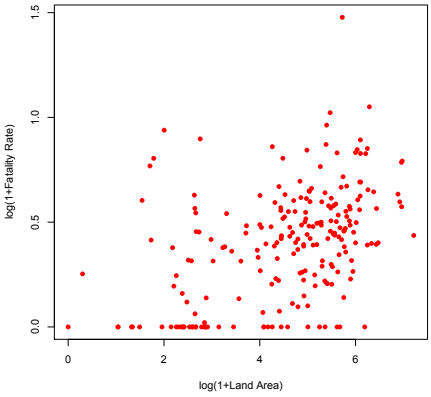



(a) Land Area vs Total Deaths




(b) Land Area vs Deaths Per Million

On the other hand, the plot of Land Area vs Deaths Per Million is not showing much linear tendency. The correlation of the logarithmic indicators is calculated to be 0.27.




Land Area vs Fatality Rate

The Pearson's coefficient calculated for the plot of Land Area vs Fatality Rate is 0.45. As the value is comparatively low we don't use regression line here.

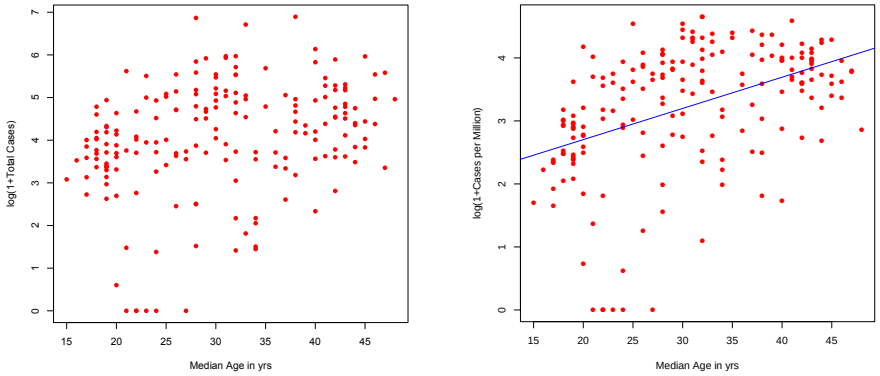
## 9.6 Median Age

We first look at the plots of Median Age vs Total Cases and Median Age vs Cases Per Million. The Pearson's correlation coefficient for the plots of Median Age vs Total Cases and vs Cases per million are 0.29 and 0.44 respectively.

Since the correlation values is not too low for Median Age vs Cases Per Million plot, and somehwat linear, we conduct the regression diagnostics. The plot passes the regression diagnostic test somewhat. We draw regression line to get a visual idea of how the data is spread around the line of best fit.

The regression parameters are calculated for these plots as follows:
For Median Age vs Cases Per Million:

$$\hat{\beta}_0 = 1.7132 \text{ and } \hat{\beta}_1 = 0.0495.$$

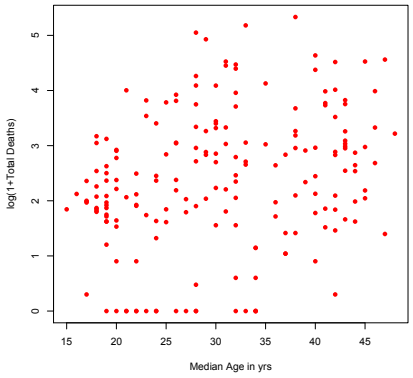



(a) Median Age vs Total Cases (b) Median Age vs Cases Per Million

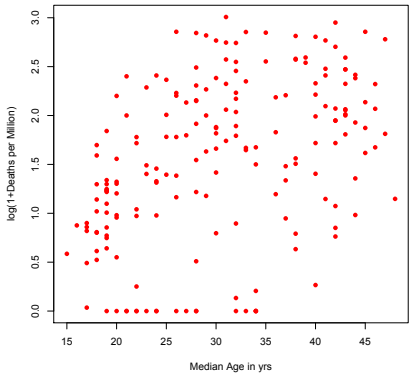
Next we look at the plots of Median Age vs Total Deaths and vs Deaths Per Million. The Pearson's correlations calculated for Median Age vs Total Deaths plot is 0.26 and for Median Age vs Deaths Per Million is 0.42.

It is interesting to note that the correlation of plots of Median Age vs Total Cases and Median Age vs Total deaths is very similar.

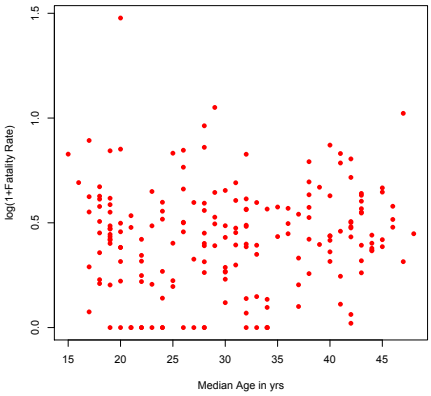
Analogously, the correlation for the plots of Median Age vs Cases per million and Median Age vs Deaths per million are very similar too.
The correlation for Median Age vs Fatality Rate plot is 0.06. Since it is so low, we do not do regression for this plot.




(a) Median Age vs Total Deaths




(b) Median Age vs Deaths Per Million




Median Age vs Fatality Rate

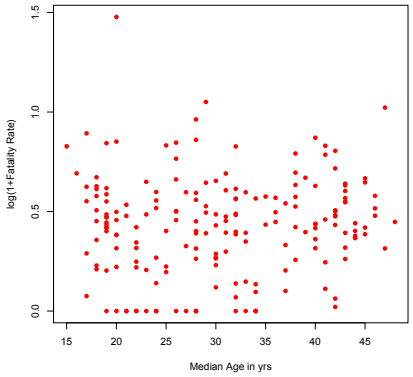
The correlation for Median Age vs Fatality Rate plot is 0.06. Since it is so low, we do not do regression for this plot.

## 9.7 Urban Population

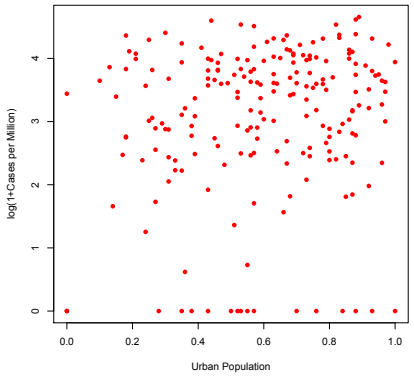
Urban Population for a country is measured in percentage values. Thus, the values are ranged from 0 to 100. So, for plotting we don't use log function for Urban Population in the plots. For the Covid indicators we use log in all the plots.

The Pearson's correlation coefficients are calculated and the values are listed bellow:

1. Plot of Urban Population vs Total cases: 0.14

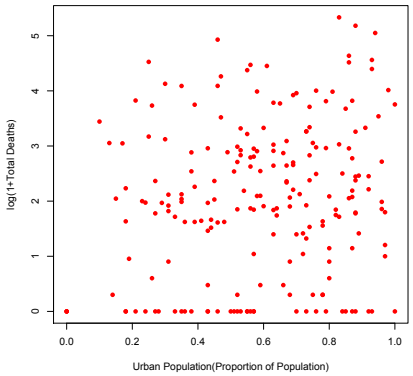



(a) Urban Pop vs Total Cases

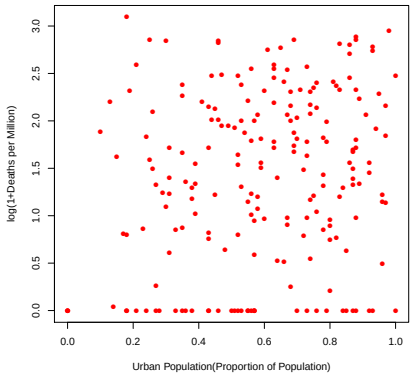



(b) Urban Pop vs Cases Per Million

2. Plot of Urban Population vs Cases Per Million: 0.11
3. Plot of Urban Population vs Total Deaths: 0.17
4. Plot of Urban Population vs Deaths Per Million: 0.09
5. Plot of Urban Population vs Fatality Rate: 0.09

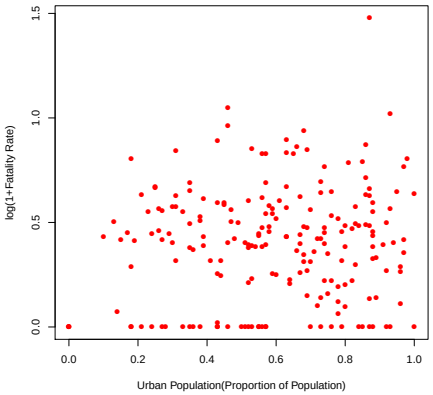



(a) Urban Pop vs Total Deaths




(b) Urban Pop vs Deaths Per Million

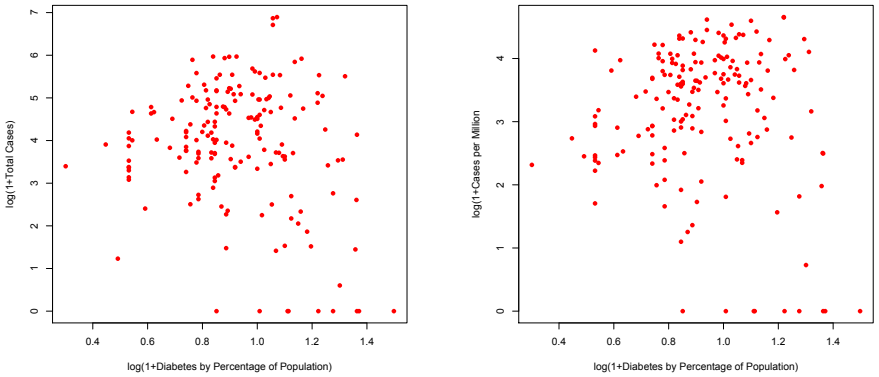
All the correlation coefficients are considerably low. The plots does not look linear and so we don't use regression lines for the above plots.




Urban Pop vs Fatality Rate

## 9.8 Diabetes Prevalence

There are several studies to show the increase in risk of communicable diseases with presence of comorbidity. Diabetes being one of the main components of comorbidity is often linked with higher risk factor. Thus we try to study it's relation with the occurrence and fatality of Covid 19. Logarithmic indicators are taken for obtaining better plots.

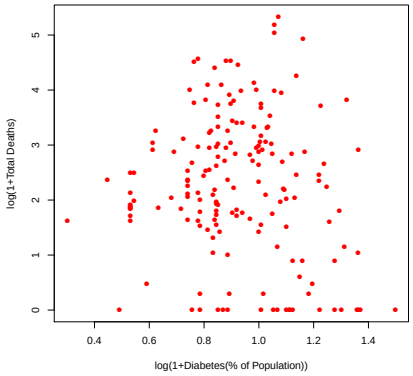



(a) Diabetes vs Total Cases (b) Diabetes vs Cases Per Million

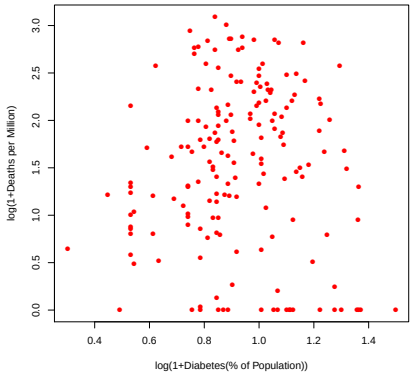
The Pearson's correlation coefficients are calculated and the values are listed bellow:

1. Plot of Diabetes vs Total cases: 0.05
2. Plot of Diabetes vs Cases Per Million: 0.10

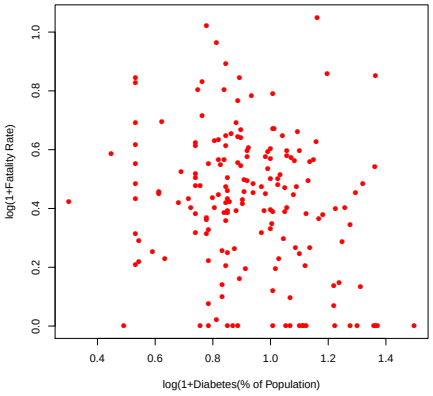
3. Plot of Diabetes vs Total Deaths: 0.04
4. Plot of Diabetes vs Deaths Per Million: -0.05
5. Plot of Diabetes vs Fatality Rate: 0.10




(a) Diabetes vs Total Deaths




(b) Diabetes vs Deaths Per Million

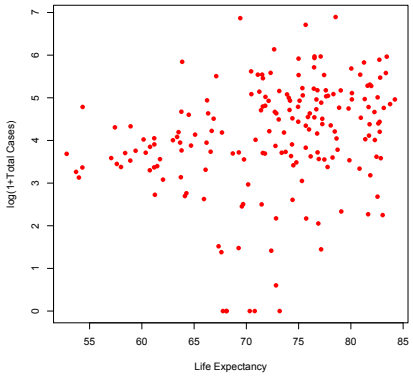



Diabetes vs Fatality Rate

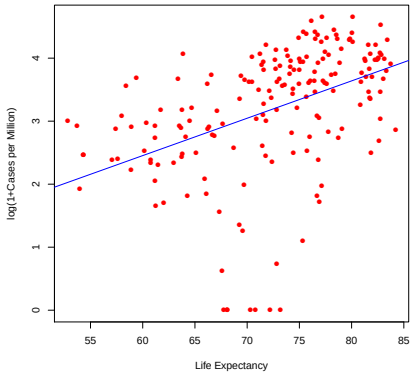
All the correlation coefficients are considerably low. Because of such low correlation values we don't use regression lines for the above plots.

## 9.9 Life Expectancy

Life expectancy is taken as the average life span over the population. As the values are not very large (almost always less than 90) we do not take log.




(a) Life Expectancy vs Total Cases




(b) Life Expectancy vs Cases Per Million

The Pearson's correlation for the plot of Life Expectancy vs Total cases is calculated to be 0.08. Because of it's low value we don't take regression line for it.

On the other hand the plot of Life Expectancy vs Cases Per Million has linear relation. The correlation coefficient is 0.42 and the plot looks linear. The plot passes regression diagnostic test. Thus, we can use regression model. The regression parameters are

$$\hat{\beta}_0 = -1.10336 \text{ and } \hat{\beta}_1 = 0.05928.$$

So, the regression line will be

$$Y = 0.05929X - 1.10336$$

where $X = LifeExpectancy$ and $Y = log(1 + CasesPerMillion)$.

The Pearson's correlation for the plot of Life Expectancy vs Total Deaths is calculated to be 0.13. We don't take regression line for it.

On the other hand the plot of Life Expectancy vs Deaths Per Million has linear relation. The correlation coefficient is 0.39. The plot passes regression diagnostic test somewhat. Thus, we can use regression model. The regression parameters are

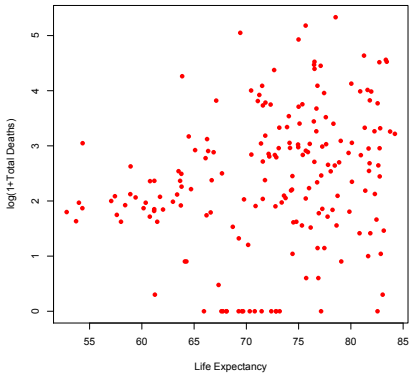
$$\hat{\beta}_0 = -1.98610 \text{ and } \hat{\beta}_1 = 0.04865.$$

So, the regression line will be

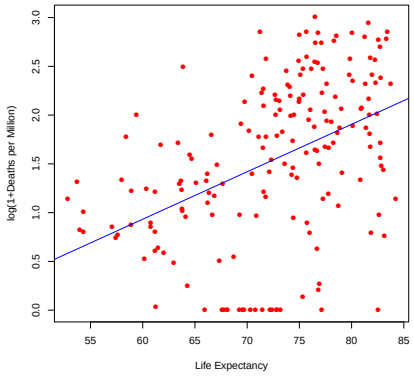
$$Y = 0.04865X - 1.98610$$

where $X = LifeExpectancy$ and $Y = log(1 + DeathsPerMillion)$.

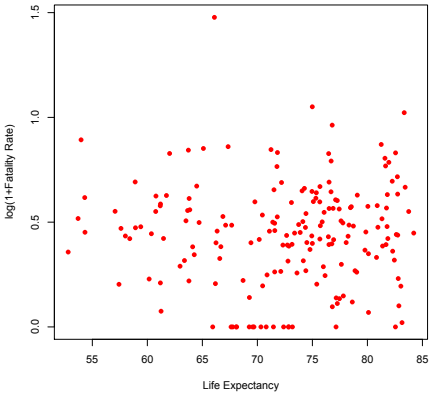
For the plot of Life Expectancy vs Fatality Rate the correlation is -0.01. The plot does not show any linear pattern, so we don't take the regression line.




(a) Life Expectancy vs Total Deaths




(b) Life Expectancy vs Deaths Per Million



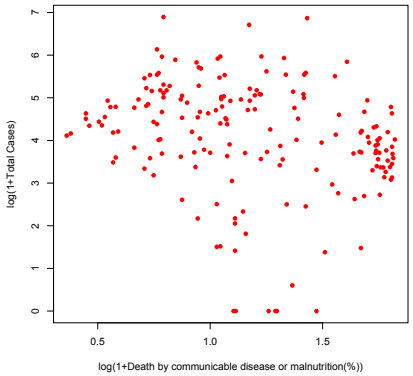

Life Expectancy vs Fatality Rate

## 9.10 Percentage of Death by Communicable diseases/ malnutrition:

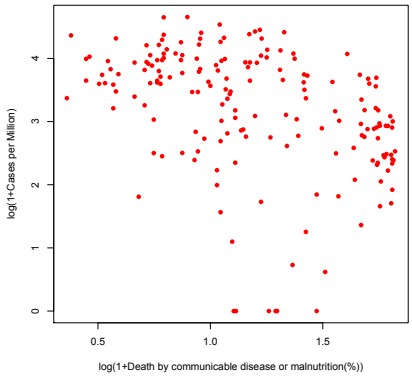
Death by communicable diseases/ malnutrition (will be referred to as DeathComm henceforth) for a country is measured as percentage of total deaths in country. Thus, the values are ranged from 0 to 100. Though the values are not very large we take $log(1 + DeathComm)$ to get better fitted plots. For the Covid indicators we use log in all the plots.

The Pearson's correlation coefficients are calculated and the values are listed bellow:

1. Plot of DeathComm vs Total cases: -0.24



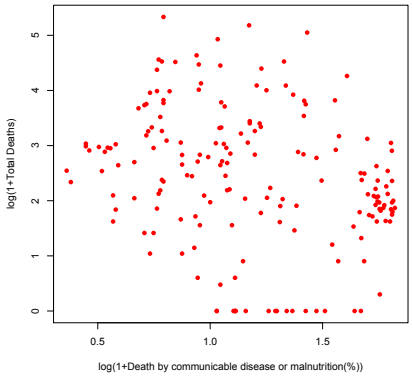

(a) DeathComm vs Total Cases



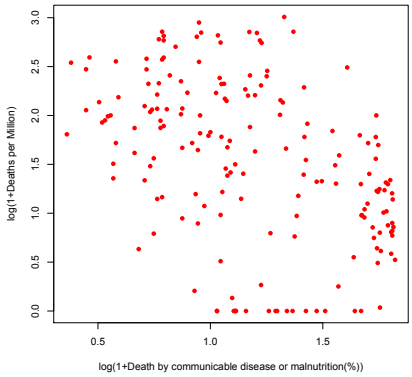

(b) DeathComm vs Cases Per Million

2. Plot of DeathComm vs Cases Per Million: -0.39
3. Plot of DeathComm vs Total Deaths: -0.26
4. Plot of DeathComm vs Deaths Per Million: -0.44
5. Plot of DeathComm vs Fatality Rate: -0.06

It is interesting to note that Deathcomm is negatively correlated with all the Covid-19 indicators.

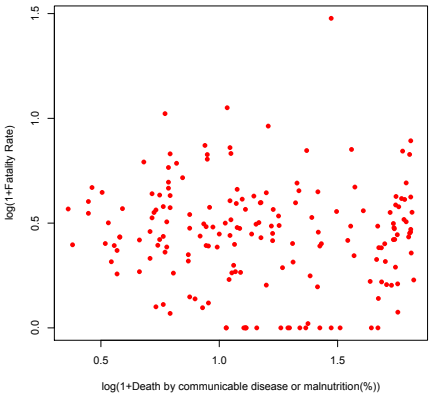



(a) DeathComm vs Total Deaths




(b) DeathComm vs Deaths Per Million

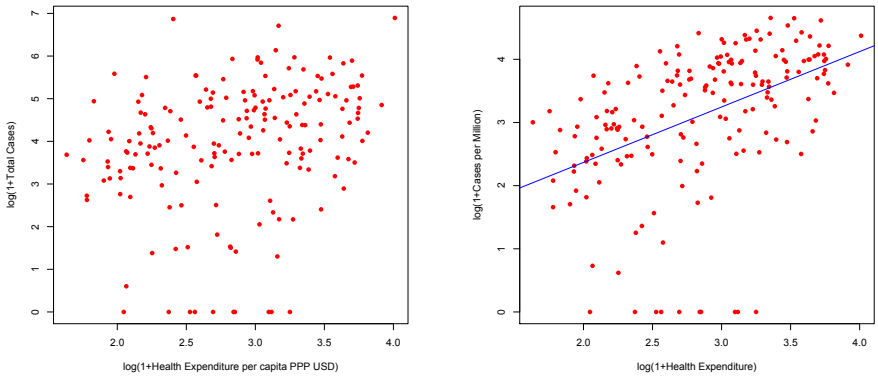
The correlation coefficient of fataliy rate is considerably low. Because of such low correlation value we don't use regression lines for the plot on Deathcomm vs Fatality Rate.




DeathComm vs Fatality Rate

## 9.11 Health Expenditure per capita PPP (USD)

Health Expenditure per capita PPP (will be referred to as HealthExp henceforth) is measured in US dollars. Firstly, we look at the plots of HealthExp vs Total Cases and vs Cases per million. The correlations for the plots of Per Capita Income vs Total Cases and vs Cases Per Million are 0.24 and 0.42 respectively.




(a) Healthexp vs Total Cases (b) Healthexp vs Cases Per Million

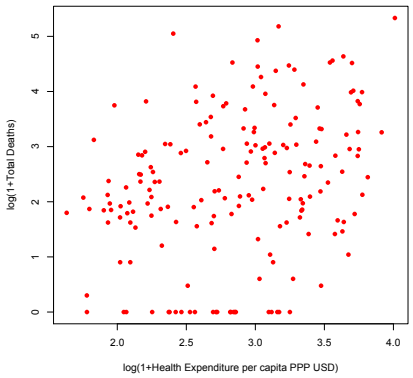
The plot passes regression diagnostic test somewhat. We plot the regression line for HealthExp vs Cases per million plot to get some idea on their relation. The regression parameters are calculated and the values are

$$\hat{\beta}_0 = 0.6141 \text{ and } \hat{\beta}_1 = 0.8770.$$

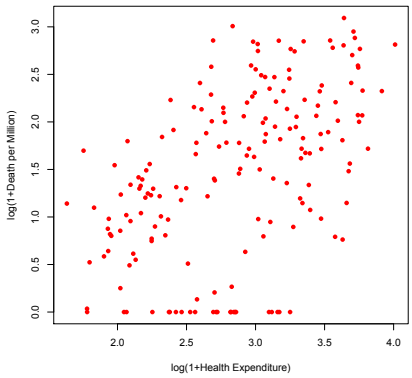
So, the regression line will be

$$Y = 0.8770X + 0.6141$$

where $X = log(1 + HealthExp)$ and $Y = log(1 + CasesPerMillion)$.

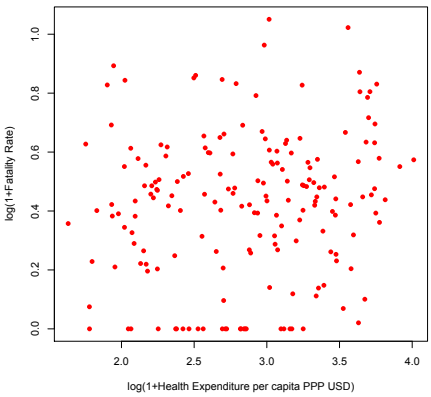



(a) HealthExp vs Total Deaths




(b) HealthExp vs Deaths Per Million

The Pearson's correlations calculated for HealthExp vs Total Deaths plot is 0.30 and for HealthExp vs Deaths Per Million is 0.41.

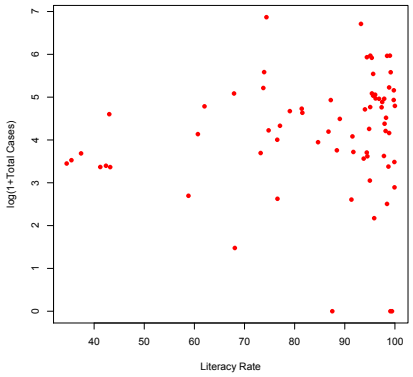



HealthExp vs Fatality Rate

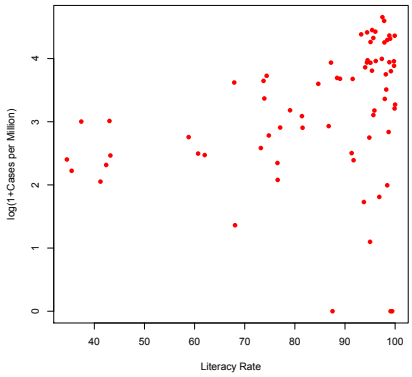
For the plot of HealthExp vs Fatality Rate the correlation coefficient is 0.13. We can't conclude any linear relation between them.

## 9.12 Literacy Rate

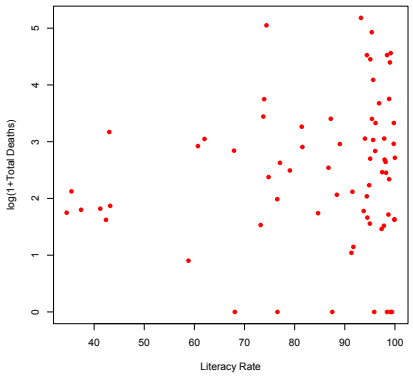
Literacy rate for a country is measured as percentage of total population in country, who are literate. Thus, the values are ranged from 0 to 100. So, for plotting we don't use log function for Literacy rates. For the Covid indicators we use log in all the plots. As Literacy Rate for many countries were not available, the plots won't be as informative as the plots of the other Country indicators.



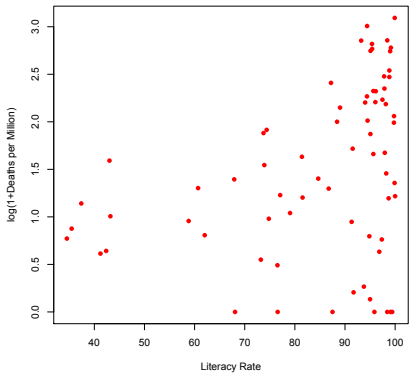

(a) Literacy vs Total Cases




(b) Literacy vs Cases Per Million




(a) Literacy vs Total Deaths




(b) Literacy vs Deaths Per Million

The Pearson's correlation coefficients are calculated and the values are listed bellow:

1. Plot of Literacy vs Total cases: 0.12
2. Plot of Literacy vs Cases Per Million: 0.30

3. Plot of Literacy vs Total Deaths: 0.13

4. Plot of Literacy vs Deaths Per Million: 0.34

5. Plot of Literacy vs Fatality Rate: -0.09

The correlation values of the plots of Literacy vs Total cases, Total deaths and Fatality Rates is low, and there is lack of linear tendency in all the plots. Hence, we do not do any regression for them.

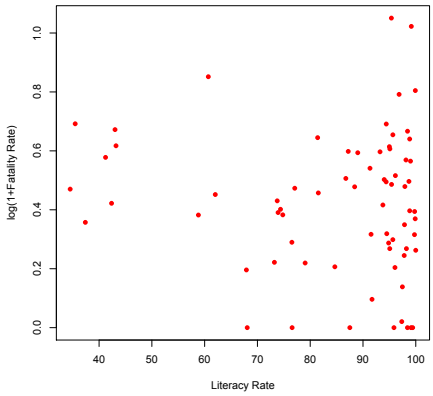
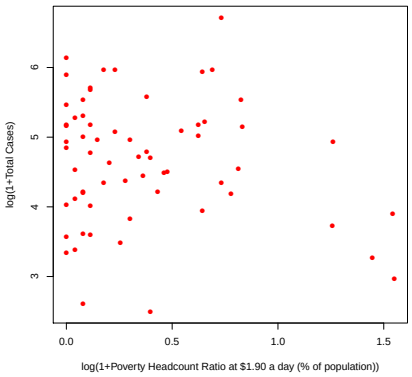
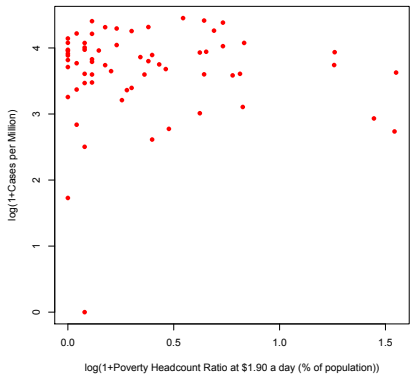



Figure 30: Literacy vs Fatality Rate

## 9.13 Poverty




(a) Poverty vs Total Cases

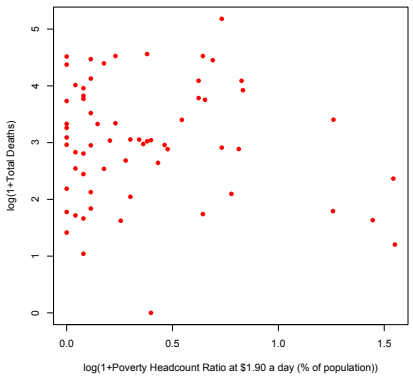



(b) Poverty vs Cases Per Million

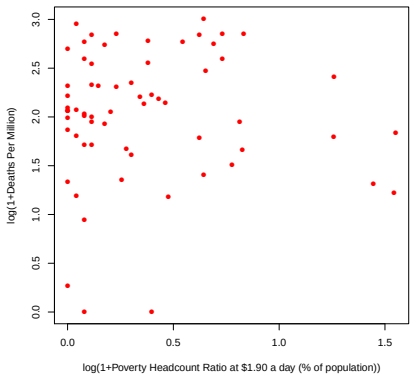
Here we measure the Poverty headcount ratio at 1.90USD a day by percentage of population. As our data set doesn't have many points having poverty values the plots won't be as informative as the plots of the other Country indicators.

The Pearson's correlation coefficients are calculated and the values are listed bellow:

1. Plot of Poverty vs Total cases: -0.04
2. Plot of Poverty vs Cases Per Million: -0.01
3. Plot of Poverty vs Total Deaths: -0.10
4. Plot of Poverty vs Deaths Per Million: -0.12
5. Plot of Poverty vs Fatality Rate: -0.02

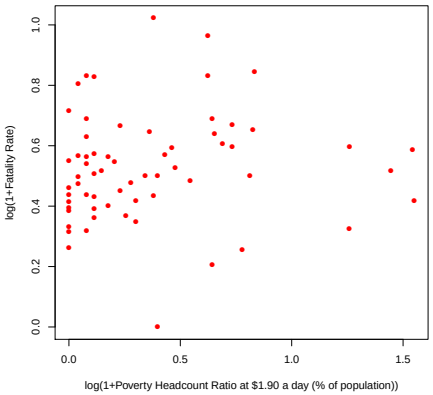



(a) Poverty vs Total Deaths




(b) Poverty vs Deaths Per Million

All the correlation coefficients are considerably low. Due to lack of linear tendency, we don't use regression lines for the above plots.




Poverty vs Fatality Rate

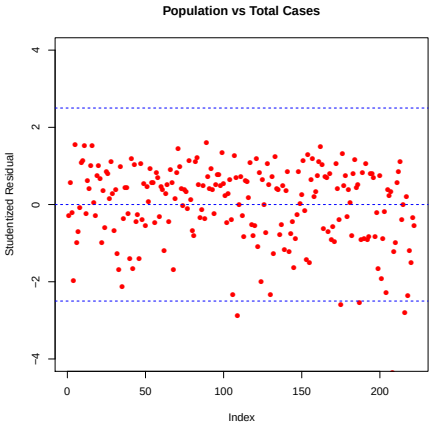
# 10 Outlier Analysis

In this section we shall be looking for outliers in the plots for which we have done regression. To check for outliers, we shall be plotting Studentized residuals against the index set. Since most of the simple linear regression models are not very high precision fit of the scatterplots, we shall be less stringent in our classification of outliers and declare any country having studentized residual greater than 2.5 in magnitude as an outlier (as opposed to the general convention of setting the threshold at 2).

## 10.1 Population vs Total Cases

The outliers as observed from the above plot are **Laos, Samao, Solomon Islands, Turkmenistan and Vanuatu**. Laos had 23 coronavirus cases and Solomon Islands had 3. All the rest have 0 officially confirmed coronavirus cases. Hence these occur as outliers. Possible explanations could include the fact that all of these countries have low median age, very less population densities, and receive very little international traffic, reducing chances of external contamination. Furthermore, countries like Solomon islands, Samoa and Vanuatu are remote islands in the Pacific, making them geographically isolated. Both Turkmenistan and Laos have socialist governments and are politically closed countries, and hence the reliability of official data as a true count of cases is low.

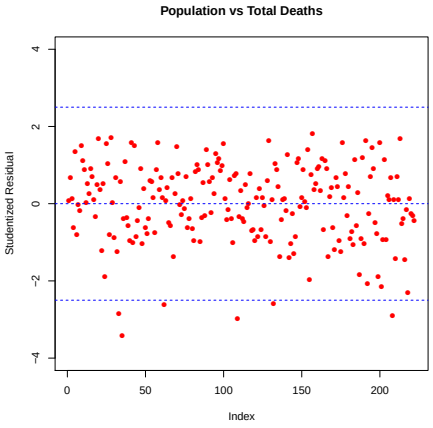
It is interesting to note that no country has a high ($> 2.5$) positive studentized residual. Thus most countries with considerable number of Covid-19 cases, have total cases reasonably close to predicted value, given their population.




Studentized Residual Plot

## 10.2 Population vs Total Deaths

The outliers as observed from the above plot are **Laos, Turkmenistan, Burundi, Cambodia, Eritrea, Mongolia**. The occurrence of Laos and Turkmenistan here are expected given that they appear as outliers (with extremely low cases) in Population vs Total Cases plot. Burundi has 1 reported death due to Covid-19, while all the other outlier countries have 0 official deaths due to Covid-19.



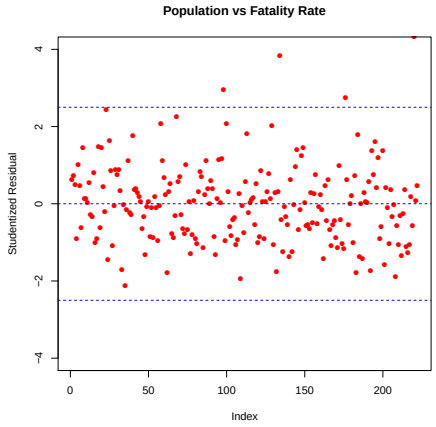

Studentized Residual Plot

An interesting thing to note that all these countries have very low median age, with Burundi and Eritrea having the least with 17 years and 19 years old

respectively. All these countries have very low total positive coronavirus cases as well, which might be attributed to the fact that they receive very low international traffic, and have low population densities. There is also less reliability of data, due to less data transparency, for these countries.

It is interesting to note that no country has a high ($> 2.5$) positive studentized residual. Thus most countries with considerable number of Covid-19 cases, have total deaths reasonably close to predicted value, given their population.

## 10.3 Population vs Fatality Rate



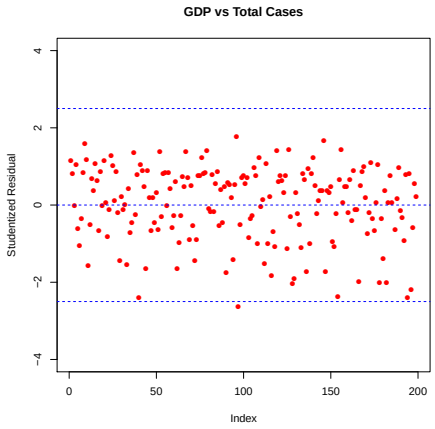

Studentized Residual Plot

The outliers as observed from the above plot are **Isle of Man, Montserrat, San Marino and Yemen**. Looking at the data, this exception can be attributed to the low number of cases in the above mentioned countries; the highest number of cases being 2057 in Yemen, and the lowest being 13 in Montserrat. Isle of Man and San Marino show 348 and 781 cases respectively, which is considerably low. Due to these low numbers, the death of even a small number of COVID19 patients contributes a lot to the proportion of deaths, and hence, the high fatality rates.

The high fatality rate in San Marino can also be related to the high population density of the country (566). Also for Yemen, it is interesting to note that the median age is considerably low - 20 years. Such low positive cases for COVID19 can be due to other reasons too such as low international traffic, and having low populations. Also there might be less reliability of data for some of these countries.

It is interesting to note that no country has a very low ($< -2.5$) negative studentized residual. Thus most countries with considerable number of Covid-19 cases, have fatality rates reasonably close to predicted value, given their population.

## 10.4 GDP vs Total Cases



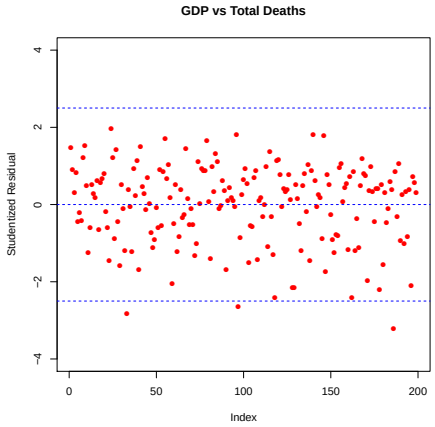

Studentized Residual Plot

The outliers as observed from the above plot are **Laos and Turkmenistan**. As cited earlier in the section of **Population vs Total Cases**, these two countries have very low COVID19 cases(Turkmenistan 0 cases, Laos 23 cases). To briefly recall the possible reasons for this, these countries have low population densities, low international traffic, slightly low median age and also they have politically closed governments, leading to a high chance of less transparent data.

It is interesting to note that no country has a high ($> 2.5$) positive studentized residual. Thus most countries with considerable number of Covid-19 cases, have total cases reasonably close to predicted value, given their population.

## 10.5 GDP vs Total Deaths

The outliers as observed from the above plot are **Laos, Cambodia and Turkmenistan**. As cited earlier in the section of **Population vs Total Cases**, both Laos and Turkmenistan have very low COVID19 cases(Turkmenistan 0 cases, Laos 23 cases), and hence even lower deaths due to COVID19. Cambodia occurs as an outlier sicne it has 0 deaths due to covid. To briefly recall the

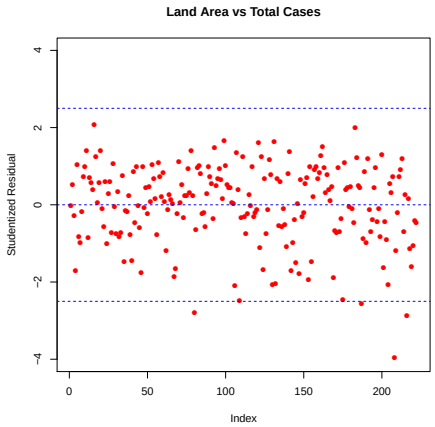



Studentized Residual Plot

possible reasons for this, these countries have low population densities, low international traffic, slightly low median age and also they have politically closed governments, leading to a high chance of less transparent data.

It is interesting to note that no country has a high ($> 2.5$) positive studentized residual. Thus most countries with considerable number of Covid-19 cases, have total deaths reasonably close to predicted value, given their population.

## 10.6 Land Area vs Total Cases

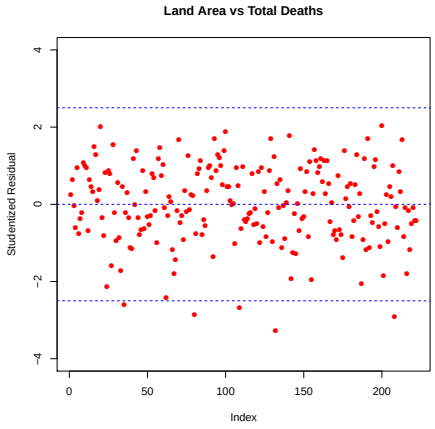



Studentized Residual Plot

The outliers observed from the above plot are **Greenland, Solomon islands, Turkmenistan** and **Vanuato**. All of these countries except Greenland, occur as outliers in Population vs Total cases plot as well. Since these countries also have low population density, it is expected that they would appear as outliers here as well.

Greenland, which is the only new country here, has a large area, with a very low population, and hence very low Covid-19 cases. These are a consequence of its hostile geographical conditions, making most of its landmass uninhabitable, and its isolated geographical location, and low international traffic.

## 10.7 Land Area vs Total Deaths

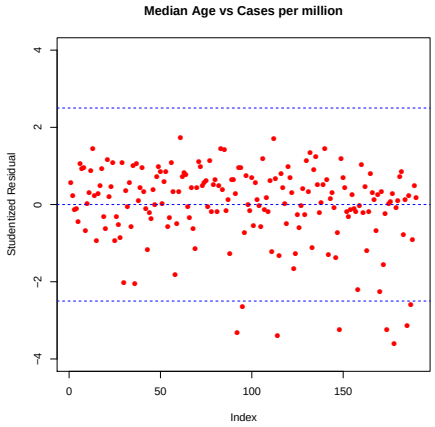



Studentized Residual Plot

The outliers observed from the above plot are **Greenland, Cambodia, Turkmenistan, Laos** and **Mongolia**. Interestingly, all these countries have 0 deaths. This is no surprise, since they have very low total cases to begin with, and all of these, except Greenland, occured as outliers in the Population vs total death plot as well. The occurence of Greenland as an outlier here is not unexpected, since it has no death, while it has a very large land area (even larger than Germany). The fact that there countries have no deaths could be explained by their low median age, as Covid-19 is mostly fatal in older people.

## 10.8 Median Age vs Cases Per Million

The outliers observed in the Studentized residual plots for this are **Kiribati, Laos, Micronesia, Samoa, Tonga, Turkmenistan, Vanuatu** and **Viet-**




Studentized Residual Plot

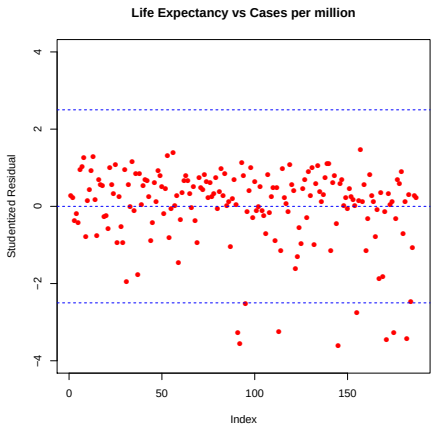
**nam**. All of these countries except Vietnam and Turkmenistan, have a median age in the range of 22-24. Vietnam has a median age of 32 and Turkmenistan has a median age of 27. However, all of these countries except Vietnam appear as outliers here, since they have 0 cases per million. It is due to this extreme value of cases per million, due to which they show significant deviation from predicted linear model. The reasons for this have been explained above.

However, Vietnam proves to be an interesting exception to the general trend. Among countries with median age 32 years, Vietnam has by far, the least Cases per million at around 11. The next highest value for cases per million at median age 32 years is reported by Grenada at 222. However, looking at all other parameters, we see that among all countries with median age 32, Vietnam has almost the least diabetes prevalence, and is also the poorest country on the Per Capita Income scale. We do not understand how these factors may influence its Covid-19 cases per million count against median age.

The most probable reason behind this unusual behaviours of Vietnam may be accredited to the strict and early lockdown and testing measures undertaken by its Communist government. Many had earlier thought it to be an overreaction from the country. However, clearly from the plots we see that their Draconian measures paid off in keeping the spread of Covid-19 in bounds, despite having one of the highest population densities amid countries with median age 32.

With respect to concerns of whether Vietnam's data can be trusted, no evidence of systematic cover up of cases has yet been found.

## 10.9 Life Expectancy vs Cases Per Million




Studentized Residual Plot

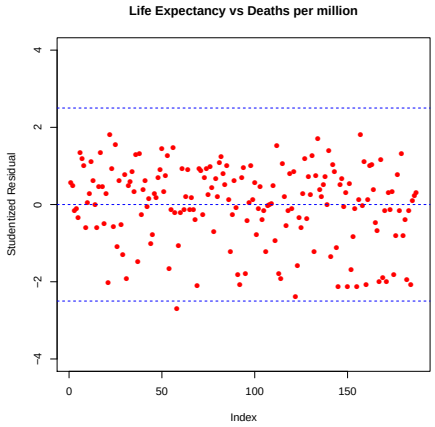
The outliers observed in the studentized residual plots for this are **Kiribati, Kosovo, Micronesia,Laos, Samoa, Solomon Islands, Tonga, Turkmenistan** and **Vanuatu**. All these countries except Kosovo have very low positive COVID19 cases reported; Kiribati,Micronesia,Samoa,Tonga,Turkmenistan have zero official cases whereas Solomon Islands has three and Laos has 23. Kosovo has 16,459 reported positive cases, however the data shows zero cases per million. Clearly, this shows that there is an error in the data.

Keeping Kosovo aside, the cause for such low cases can be attributed to the considerably low population densities in all these countries, the highest being 164 in Micronesia and lowest being 13 in Turkmenistan. However, the official data provided by some of these countries might be inaccurate due to politically closed governments. Also, geographically, Kiribati, Micronesia, Samoa, Solomon Islands, Tonga and Vanuatu are isolated islands, which can also be a cause for low positive COVID19 cases. Another reason can be the low international traffic in most of these countries.It is interesting to note that no country has a high ($> 2.5$) positive standardized residual. Thus most countries with considerable number of Covid-19 cases, have cases per million reasonably close to predicted value, given their population.

## 10.10 Life Expectation vs Deaths Per Million

The outlier observed in the studentized residual plots for this is only **Faeore Islands**.

We see that Faeroe Islands have 0 deaths per million, since they have 0



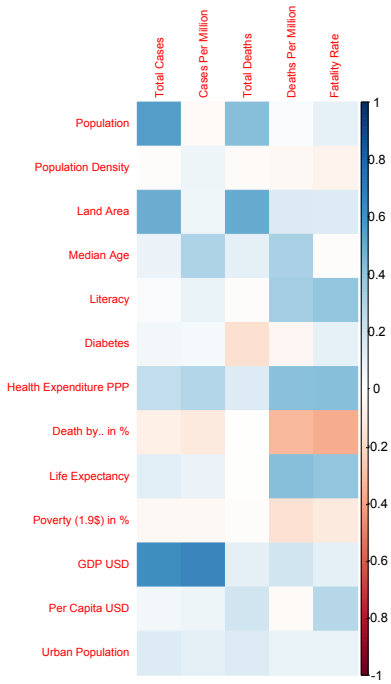

Studentized Residual Plot

total deaths. Among all countries / territories with 0 deaths per million, Faeroe islands have the highest life expectancy at around 82 years. However we cannot explain why it behaves as an outlier, since we do not have sufficient data for other country-wise indicators for Faeroe island.

# 11 Correlation Plots

In this section we use correlation plots to visualize the correlation values for each Country indicator vs Covid indicator. The correlation values are calculated by the Pearson's correlation coefficient.

We first take the correlation plot for the Country indicator vs Covid indicator directly, without taking any function of the indicators. In the rows we have 5 Covid indicators and in the columns we have all the 13 Country indicators. For each Country indicator we enter the correlation values against each of the 5 Covid indicators.



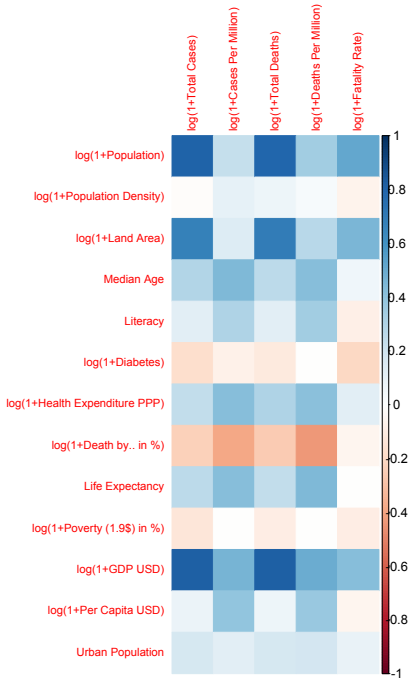

Correlation Plot

Most of the Country indicators doesn't show much linear tendency with the Covid parameters. Thus, the above correlation plot is mostly containing light shades. Note that light shades indicates small correlation values (close to 0).
On the other hand, in most of the cases taking logarithmic functions of both the Country indicators and Covid indicators give more linear plots. However in some cases we take Country indicator and log of Covid indicator to get a better fitted plot. The plots which were studied in Section 9 were carefully chosen to be the best fitted plots.
The correlation plot taking the particular functions of the indicators into account is given bellow. We get much darker shades than the previous plot in the next one.

We can list a few observations here-

1. Prominent positive relations with Covid parameters are noted for Population, Land Area, Median Age, Health Expenditure, GDP Per Capita and GDP.

2. Death by Communicable Diseases, Diabetes Prevalence and Poverty is showing negative relation with the Covid indicators. However except Death by Communicable disease all the correlations are small negative values.




Correlation Plot for appropriate functions

3. Covid Total Cases and Total Deaths are behaving very similarly with the Country indicators.

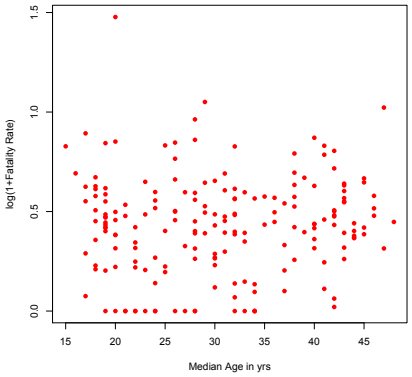
4. Whenever Covid Prevalence is having high correlation with some Country indicator, Incidence is having small correlation. For example Population, Land Area and GDP are highly correlated with Total Cases and Total Deaths but are having small correlation with Cases Per Million and Deaths Per Million.

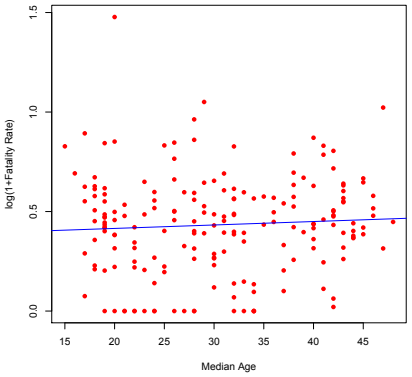
# 12 Median Age vs Fatality Rate analysis

In this section we explore the relation between Median Age and Fatality Rate. Our main aim will be to see if we can show that they are not related.

We first take a look into their scatter plot. The points in scatter plot is well distributed and also it shows linear tendency. We conduct the regression diagnostics for the plot of $MedianAge$ vs $log(FatalityRate)$.

From regression diagnostic plots it is evident that the model assumptions hold. The qq plot for normality is also close to a straight line and moreover we have our sample almost same as the population.




The regression parameters are

$$\hat{\beta}_0 = 0.3804 \text{ and } \hat{\beta}_1 = 0.0017.$$

The Pearson's Correlation coefficient for the plot is 0.06.

## 12.1 $\beta_1 = 0$ Testing

Null Hypothesis: $\beta_1 = 0$.

Alternative Hypothesis: $\beta_1 \neq 0$

The test statistic value is coming to be $t = 0.853176$. The rejection region for $\alpha = .05$ is

$$|t| > t_{\alpha/2} = 1.972.$$

So, as $|t| < t_{\alpha/2}$ we do not reject the null hypothesis. Hence, there is not enough evidence to say that Median Age can be used to predict something about the Fatality Rate. This means that the deterministic part of Fatality Rate (expected value at a given Median Age) does not change as the Median Age changes.

## 12.2 Chi-Square Testing

The Chi Squared test for independence of classification data will be used. The data will be classified using the variables, Median Age and Fatality Rate in the following manner :

1. The units are firstly ranked according to both Median Age and Fatality Rate separately.
2. They then are classified into discrete intervals of percentiles obtained by this ranking.

3. They are then crossed across the classifications of the two variables.

4. This information is then written into a table where the columns are the interval classifications of Median Age and the rows are the interval classifications of Fatality Rate.

5. So, each cell of the table will holds the number of units that belong to the intervals of classification of the two variables which correspond to the row and column of the cell.

| 2 way - Categorical Table | 20yrs & below | 21-27 years | 28-32 years | 33-40 years | 40yrs & above |
|---|---|---|---|---|---|
| 0.000%-0.674% | 6 | 14 | 6 | 9 | 3 |
| 0.674%-1.492% | 6 | 5 | 11 | 7 | 9 |
| 1.492%-2.062% | 10 | 7 | 6 | 6 | 9 |
| 2.062%-3.113% | 7 | 7 | 8 | 8 | 8 |
| 3.113% & above | 10 | 5 | 8 | 6 | 9 |

Two Way Categorical Table

| 2 way - Categorical Table | 20yrs & below | 21-27 years | 28-32 years | 33-40 years | 40yrs & above |
|---|---|---|---|---|---|
| 0.000%-0.674% | 7.8 | 7.6 | 7.8 | 7.2 | 7.6 |
| 0.674%-1.492% | 7.8 | 7.6 | 7.8 | 7.2 | 7.6 |
| 1.492%-2.062% | 7.8 | 7.6 | 7.8 | 7.2 | 7.6 |
| 2.062%-3.113% | 7.8 | 7.6 | 7.8 | 7.2 | 7.6 |
| 3.113% & above | 7.8 | 7.6 | 7.8 | 7.2 | 7.6 |

Two Way Categorical Table - Expected

The hypothesis testing for independence is now done :

Null Hypothesis is that the two classifications are independent and Alternate Hypothesis that the two classifications are dependent

Test Statistic : 16.094.

Rejection Region by the Chi Squared Distribution (at $\alpha = 0.05$ and 16 df) : $[26.29, \infty)$

Null Hypothesis is not rejected. So we can say with 95 certainty that Median Age and Fatality Rate, as described by the categories may be independent.

# 13 Review of some Plots excluding Zero Deaths

In few plots, the regression line might be significantly affected if there are many countries having 0 deaths/cases. Here, we avoid the countries with 0 deaths in some plots and then draw the regression lines for them and identify outliers.

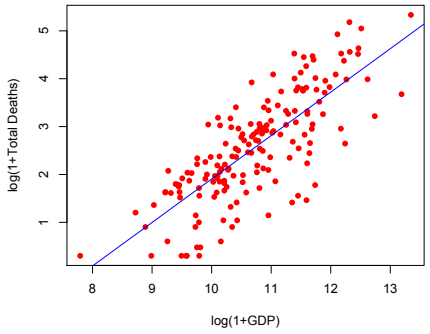
## 13.1 Land Area vs Total Deaths

### 13.1.1 Regression Model

All the countries having positive number of deaths due to Covid is plotted with their Land Area. The regression parameters are

$$\hat{\beta}_1 = 0.5842 \text{ and } \hat{\beta}_0 = -0.3201$$

The correlation value is 0.6610.

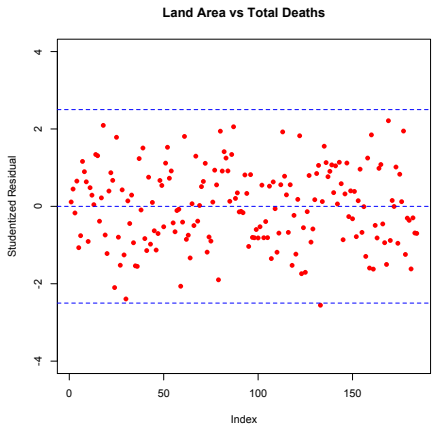



Land Area vs Total Deaths

### 13.1.2 Outliers

Two outliers were observed, namely **Papua** and **New Guinea**, as compared to the five observed when zero deaths countries were included (**Greenland, Cambodia, Turkmenistan, Laos, Mongolia**. This shows that the zero

death countries influenced the regression line significantly. Also, as the countries with which this linear model was constructed have nonzero deaths, the outliers observed now help us identify exceptional behaviour among countries with nonzero deaths, which is more beneficial.




Land Area vs Total Deaths

## 13.2 GDP vs Total Deaths

### 13.2.1 Regression Model

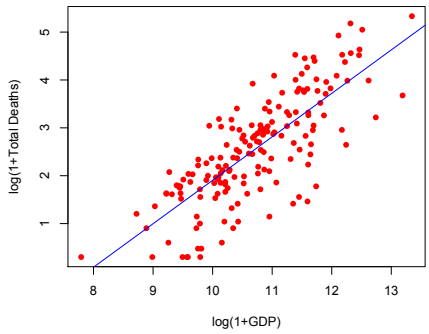
All the countries having positive number of deaths due to Covid is plotted with their GDP. The regression parameters are

$$\hat{\beta_1} = 0.9077 \text{ and } \hat{\beta_0} = -7.1726$$

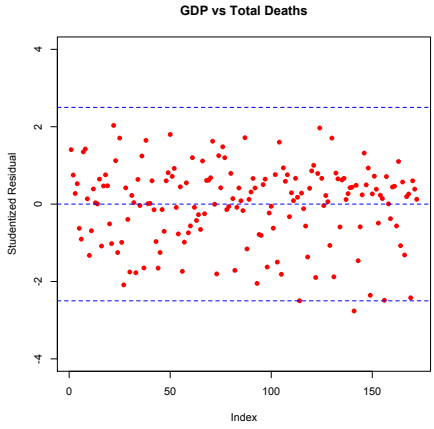
The correlation value is 0.7765.

### 13.2.2 Outliers

One outlier was observed, namely **Singapore**, as compared to the three observed when zero deaths countries were included (**Cambodia, Turkmenistan, Laos**). This shows that the zero death countries influenced the regression line significantly. Also, as the countries with which this linear model was constructed have nonzero deaths, the outliers observed now help us identify exceptional behaviour among countries with nonzero deaths, which is more beneficial.




GDP vs Total Deaths




GDP vs Total Deaths

## 13.3 Population vs Total Deaths

### 13.3.1 Regression Model

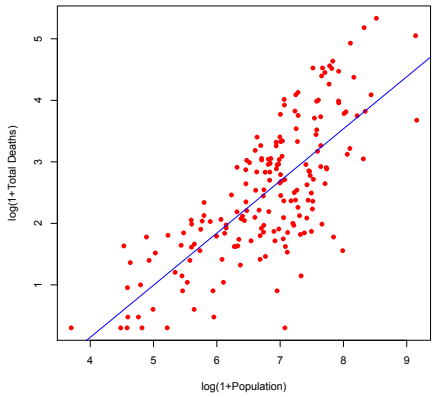
All the countries having positive number of deaths due to Covid is plotted with their Population. The regression parameters are

$$\hat{\beta}_1 = 0.8472 \text{ and } \hat{\beta}_0 = -3.2415$$

The correlation value is 0.7453.

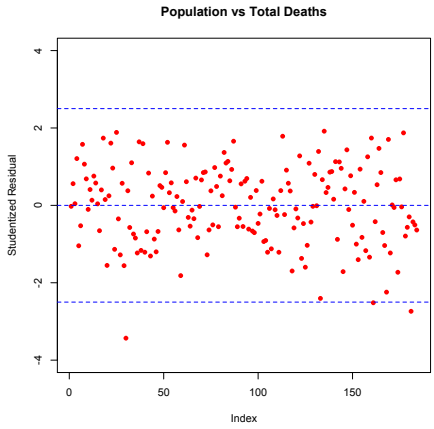
### 13.3.2 Outliers

Three outliers were observed, namely **Burundi,Sri Lanka** and **Vietnam**, as compared to the six observed when zero deaths countries were included (**Laos,**




Pop vs Total Deaths

**Turkmenistan, Burundi, Cambodia, Eritrea, Mongolia**. This shows that the zero death countries influenced the regression line significantly. Also, as the countries with which this linear model was constructed have nonzero deaths, the outliers observed now help us identify exceptional behaviour among countries with nonzero deaths, which is more beneficial.




Population vs Total Deaths

Comparing the regression coefficients of the above plots with the actual plots we can observe that the $\hat{\beta_0}$ is increasing when we avoid the countries with 0 Covid Deaths. However, the $\hat{\beta_1}$ is slightly decreased.

# 14 Synopsis

In this section we summarize all the results and observations we get by analyzing the spatial distribution of Covid 19 incidence and prevalence.

1. Covid Total Cases and Covid Total Deaths is showing positive relation with Country variables like Population, GDP, and Land Area.
2. Covid Cases Per Million and Covid Deaths Per Million are not showing any prominent relation with the Country indicators we tested.
3. Fatality Rate is having a positive relation with Population. However, the linear model is not very strong.
4. In all the scatter plots, except for the one for Population vs Fatality Rate, we observed that the countries with data way lower than the predicted values occurred as outliers. This means that countries with considerable point prevalence and incidence for total cases and total deaths fit well in our model.
5. However, in Population v/s Fatality Rate, the outliers observed show an exceptionally high value, because of which they appear in the list. There is no linear relationship between Median Age and Fatality Rate. The expected slope of the linear regression line is 0.
6. 5 classifications are made using Median Age and Fatality Rate percentiles. By Chi Square Testing for independence we show that the classifications are independent at 95 percent level.

# 15 References


1. "Spatial Analysis of Global Variability in Covid-19 Burden"- by Miller et. al.
2. "Data regarding country-specific variability in Covid-19 prevalence, incidence, and case fatality rate" - by Miller et. al.
3. Introduction to regression in R by UCLA IDRE Statistical consulting group
4. Regression diagnostics notes McMaster University Canada by John Fox
5. Queen Mary University of London Simple linear regression notes